



# **PROCEEDINGS OF**

## **2020 5<sup>TH</sup> INTERNATIONAL CONFERENCE ON GREEN TECHNOLOGY AND SUSTAINABLE DEVELOPMENT (GTSD)**

**27-28 NOVEMBER 2020 - HO CHI MINH CITY, VIETNAM**

**IEEE XPLORE VERSION**  
**(FOR STORAGE ONLY)**

# PROCEEDINGS OF

2020 5<sup>TH</sup> INTERNATIONAL CONFERENCE ON GREEN  
TECHNOLOGY AND SUSTAINABLE DEVELOPMENT (GTSD)

**(FOR STORAGE ONLY, NOT FOR SALE)**

Virtual Conference, November 27-28, 2020

Ho Chi Minh City, Vietnam

Co-organized by:



In cooperation with:



Sponsored by:



Springer

IEEE  
**SMC**

Systems, Man, and Cybernetics Society

 **IEEE**



# TABLE OF CONTENT

Copyright page	ix
Organizing Committee information	x
Authors index	xi
Numerical Study on Vibration Control of Structures Using Multi Tuned Liquid Dampers with High Mass Ratio <i>Bui Pham Duc Tuong, Phan Duc Huynh</i>	1
Modelling the Distribution of TiC Reinforced Co-Based on Skd61 Steel Surface Prepared by Laser Cladding <i>Ngoc Thien Tran, Hong Nga Thi Pham</i>	5
Real-time Measurement and Prediction of Ball Trajectory for Ping-pong Robot <i>Vo Duy Cong, Le Duc Hanh, Le Hoai Phuong</i>	9
Effect of the Welding Parameters On Mechanical Properties of AA5083 friction stir welding <i>Phan Thanh Nhan</i>	15
An improvement of Disk Aware Discord Discovery Algorithm for Discovering Time Series Discord <i>Nguyen Thanh Son</i>	19
An Adaptive Fuzzy for PMSM to Overcome the Changing Load <i>Nguyen Vu Quynh</i>	24
Crack Patterns in Direct Tension and Flexure of Ultra-High-Performance Fiber-Reinforced Concrete With Scale Effect <i>Duy-Liem Nguyen, Vu-Tu Tran, Huynh-Tan-Tai Nguyen, Min-Kyoung Kim</i>	30
Indirect Tensile Strengths of Crushed-Sand Concretes in Correlation with Its Compressive Strength <i>Duy-Liem Nguyen, Vu-Tu Tran, Thi-Ngoc-Han Vuong, Tien-Tho Do</i>	36
A method to analyzing and clustering aggregate customer load profiles based on PCA <i>Alessandro Bosisio, Alberto Berizzi, Andrea Vicario, Andrea Morotti, Bartolomeo Greco, Gaetano Iannarelli, Dinh-Duong Le</i>	41
The Bearing Capacity of Compacted Clay Reinforced by Geotextile and Sand Cushion <i>Minh-Duc Nguyen</i>	48
An Improvement of Maximum Power Point Tracking Algorithm Based on Particle Swarm Optimization Method for Photovoltaic System <i>Xuan Truong Luong, Van Hien Bui, Duc Tri Do, Thanh Hai Quach, Viet Anh Truong</i>	53
Static Analysis of Sandwich Plates using ES-MITC3 Elements based on the Third-order Shear Deformation Layerwise Theory <i>Thanh Chau-Dinh, Loi Dang-Huu, Jin-Gyun Kim</i>	59
P2C2-Popular Content Prediction and Collaboration in mobile edge caching <i>Dung Ong Mau, Anh Phan Tuan, Tuyen Dinh Quang, Quyen Le Ly Quyen, Nga Vu Thi Hong</i>	66
Ritz solution for static analysis of thin-walled laminated composite I-beams based on first-order beam theory <i>Ngoc-Duong Nguyen, Trung-Kien Nguyen</i>	73

A Redundant Unit Form of Quasi-Z-source T-Type Inverter with Fault-Tolerant Capability <i>Duc-Tri Do, Vinh-Thanh Tran, Hieu-Giang Le, Thanh-Hai Quach, Viet-Anh Truong, Minh-Khai Nguyen, Thi-Ngoc-Han Vuong</i>	78
A High-Efficient Power Converter for Thermoelectric Energy Harvesting <i>Van-Khoa Pham</i>	82
Collaborative Robotics in Construction: A Test Case on Screwing Gypsum Boards on Ceiling <i>Milan Gautam, Hannu Fagerlund, Blerand Greicevci, Francois Christophe, Jarmo Havula</i>	88
Impact of Financial Inclusion on Economic Growth: GMM Approach <i>Khac Hieu Nguyen, Thi Anh Van Nguyen</i>	94
Ultimate Bond Strength of Steel Bar Embedded in Sea Sand Concrete under Different Curing Environments <i>Quoc Khanh Tran, Tri Thuong Ngo, Duy Liem Nguyen, Ngoc Thanh Tran</i>	98
Performance Analysis and Evaluation of Underlay Two-Way Cooperative Networks with NOMA <i>Nguyen Duc Anh, Pham Ngoc Son</i>	103
Reusing Fabric Scraps in Garment Industry - A Green Manufacturing Process <i>Thi Bich Dung Phung, Tuan-Anh Nguyen</i>	109
Analyzing Total Quality Management of Service Enterprises in Vietnam <i>Nguyen Thi Anh Van, Nguyen Khac Hieu</i>	114
Synthesis of Zinc Oxide Nanoparticles and Their Antibacterial Activity <i>Thi Duy Hanh Le, Khanh Son Trinh</i>	119
A Study on Design and Fabrication of Concrete Pipe Cutting Machine <i>Cong Binh Phan</i>	124
Optimal Day-ahead Energy Scheduling of Battery in Distribution Systems Considering Uncertainty <i>Ying-Yi Hong, Man-Yin Wu, Sheng-Huei Lee</i>	128
Improving the Adaptivities of Over-The-Top Television System <i>Ha Tran Thu, Son Tran Minh, Thai Nguyen Van, Minh Le Hoang</i>	133
Multi-class Support Vector Machine Algorithm for Heart Disease Classification <i>Thanh-Nghia Nguyen, Thanh-Hai Nguyen, Duc-Dung Vo, Truong-Duy Nguyen</i>	137
Computational Intelligence Towards Trusted Cloudlet Based Fog Computing <i>Thinh Vinh Le, Tran Thien Huan</i>	141
Traditional Method Meets Deep Learning in an Adaptive Lane and Obstacle Detection System <i>Van-Tin Luu, Viet-Cuong Huynh, Vu-Hoang Tran, Trung-Hieu Nguyen, Thi-Ngoc-Hieu Phu</i>	148
A Lightweight Model For Real-time Traffic Sign Recognition <i>Trung-Hieu Nguyen, Vu-Hoang Tran, Van-Dung Do, Van-Thuyen Ngo, Thanh-Thanh Ngo-Quang</i>	153
Design of Delta Robot Arm based on Topology Optimization and Generative Design Method <i>Thanh Hai Tuan Tran, Dinh Son Nguyen, Nhu Thanh Vo, Hoai Nam Le</i>	157
Effects of Soaking Process on CBR Behavior of Geotextile Reinforced Clay with Sand Cushion <i>Tu Nguyen Thanh, Duc Nguyen Minh</i>	162

A Strategy to Enhance Generator Efficiency of Sudoku-based PV Arrays Under Partial Shading Conditions <i>Le Viet Thinh, Nguyen Duc Tuyen, Vu Xuan Son Huu</i>	168
The Influence of Water Content and Compaction on the Unconfined Compression Strength of Cement Treated Clay <i>Minh-Duc Nguyen, Anh-Thang Le, Thien-An Nguyen, Nguyen-Thao Thach, Thanh-Kiet Phan</i>	175
Traffic Flow Estimation Using Deep Learning <i>Tran Nhat Huy, Bui Ha Duc</i>	180
A Finite-Time Robust Control for a Manipulator with Output Constraints and Unknown Control Directions <i>Duc Thien Tran, Manh Son Tran, Nguyen Van Hiep</i>	185
Determine the percentage of Recovering FCC Spent Catalysts as mineral filler in the asphaltic concrete mixture <i>Anh Thang Le, Nguyen Manh Tuan</i>	192
The Effect Of Capital Structure On Profitability: An Empirical Analysis of Vietnamese Listed Banks <i>Tran Thuy Ai Phuong, Nguyen Thi Anh Van, Nguyen Thi Hoang Anh</i>	198
Carboxymethyl Cellulose /Aloe Vera Gel Edible Films For Food Preservation <i>Hung Ngoc Nguyen, Khoe Dang Dinh, Linh T. K. Vu</i>	203
The Load-Bearing of Concrete Beams as the Steel Reinforcements Connected by the Coupler at a Cross-Section of a Beam <i>Thanh-Hung Nguyen, Phuong-Doanh Huynh, Anh-Thang Le</i>	209
Optimum of Biodegradable Plate Production from Banana Trunk Waste by Taguchi Methods <i>Nhung Thi-Tuyet Hoang, Anh Thi-Kim Tran, Phan Thi Thu Thuy</i>	214
Wheelchair Navigation System using EEG Signal and 2D Map for Disabled and Elderly People <i>Ba-Viet Ngo, Thanh-Hai Nguyen, Van-Thuyen Ngo, Dang-Khoa Tran, Truong-Duy Nguyen</i>	219
Experimental study of the Strain Localization in a Rock Analogue Material at Brittle-Ductile Transition <i>Thi-Phuong-Huyen Tran, Sy-Hung Nguyen, Stéphane Bouissou</i>	224
Effects of the Relative Humidity on the Performance of Thermoelectric Freshwater Generator using Solar Power Source <i>Le Minh Nhut, Dang Thi Truc Linh</i>	232
Effect of Alkaline Solution and Curing Conditions on the Strength of Alkali – Activated Slag Mortar <i>Tai Tran Thanh, Tu Nguyen Thanh, Hyug-Moon Kwon</i>	236
Design and Control of a 4-bar-transmission 2-DOF Robot <i>Pham Tan Phat, Bui Manh Huy, Tran Hoai Nam, Huynh Vinh Nghi, Dang Xuan Ba</i>	241
The Morphological Characteristics and Physical Properties of Porous Corn Starch Hydrolyzed by Mixture of $\alpha$ -Amylase and Glucoamylase <i>Dang My Duyen Nguyen, Thanh Tung Pham</i>	247
PyPSA-VN: An open model of the Vietnamese Electricity System <i>Markus Schlott, Bruno Schyska, Dinh Thanh Viet, Vo Van Phuong, Duong Minh Quan, Ma Phuoc Khanh, Fabian Hofmann, Lueder von Bremen, Detlev Heinemann, Alexander Kies</i>	253
Customer satisfaction with service quality: An empirical study of An Giang Power Company <i>Hong-Xuyen Ho Thi, Lan Anh Nguyen Thi</i>	259

Engineering Properties of Cement Mortar Produced with Mine Tailing as Fine Aggregate <i>Duy-Hai Vo, Khanh-Dung Tran Thi, Mitiku Damtie Yehualaw, Chao-Lung Hwang, Thi-My Ngo, Hoang-Anh Nguyen</i>	264
Effect of Water-To-Solid Ratio on the Strength Development and Cracking Performance of Alkali-Activated Fine Slag under Water Curing Condition <i>Duy-Hai Vo, Khanh-Dung Tran Thi, Mitiku Damtie Yehualaw, Chao-Lung Hwang, Hoang-Anh Nguyen, Vu-An Tran</i>	268
Determining Optimal Location and Sizing of STATCOM Based on PSO Algorithm and Designing Its Online ANFIS Controller for Power System Voltage Stability Enhancement <i>Huu Vinh Nguyen, Hung Nguyen, Kim Hung Le, Minh Tien Cao, Tan Hung Nguyen, Tien Hoang Nguyen, Minh Vuong Le</i>	272
An MCS-based Model to Qualify the Relationship between Worker's Experience and Construction Productivity <i>Duy-Khanh Ha, Soo-Yong Kim, Van-Khoa Nguyen</i>	280
Chances and Challenges of Vietnam's Garment Industry in the new Trend of Sustainable Development <i>Tri Tran Quang, Tu Tran, Alang Tho, John Burgess</i>	286
Extraction of Pectin from <i>Passiflora edulis</i> by Aqueous Two-Phase System <i>Nga Thi Vo, Thi Hao Cao, Minh Hao Hoang</i>	291
Unknown Input Based Observer Design for Wind Energy Conversion System with Time-Delay <i>Van-Phong Vu, Wen-June Wang, Van-Thuyen Ngo, Dinh-Nhon Truong, Pei-Jun Lee, Ton Duc Do and Van-Ngoc Tong</i>	296
Analysis of Flexible Pavements Comprised of Conventional and High Modulus Asphalt Concrete Subjected to Moving Loading using Linear Viscoelastic Theory <i>H.T. Tai Nguyen, Thanh-Nhan Phan, Tien-Tho Do, Duy-Liem Nguyen, Vu-Tu Tran</i>	302
Effects of Forta-fi Fiber on the Resistance to Fatigue of Conventional Asphalt Mixtures <i>Tien-Tho Do, Duy-Liem Nguyen, Vu-Tu Tran, H.T. Tai Nguyen</i>	312
Application of Element Combine39 to Reflect the Nature of Newly Puzzel Shaped Crestbond Rib Shear Connector in Composite Beam <i>Pham Duc Thien, Dao Duy Kien, Nguyen Van Khoa, Nguyen Thanh Hung</i>	317
A Mobile Deep Convolutional Neural Network Combined with Grad-CAM Visual Explanations for Real Time Tomato Quality Classification System <i>Loc-Phat Truong, Bach-Duong Pham, Quang-Huy Vu</i>	321
Design of Driver Circuit to Control Induction Motor Applied in Electric Motorcycles <i>Dinh Cao Tri, Le Thanh Phuc</i>	326
Fractional order Modeling and Control of a Quadruple-tank Process <i>Lam Chuong Vo, Luan Vu Truong Nguyen, Moonyong Lee</i>	334
Treatment of Wastewater Containing Reactive Dyes by electro-Fenton Method <i>Nguyen Thai Anh, Tran Tien Khoi, Nguyen Nhat Huy, Hoang Thi Ngoc Mai, Nguyen Hong Ngoc Linh</i>	340
Breast Image Segmentation for evaluation of Cancer Disease <i>Thanh-Tam Nguyen, Thanh-Hai Nguyen, Ba-Viet Ngo, Duc-Dung Vo</i>	344

Vision-based People Counting for Attendance Monitoring System <i>Manh Cuong Le, My-Ha Le, Minh-Thien Duong</i>	349
Electricity Demand Forecasting for Smart Grid Based on Deep Learning Approach <i>Van-Binh Nguyen, Minh-Thien Duong, My-Ha Le</i>	353
A Novel Technique for Increasing Concentration Ratio and Uniformity of Fresnel Lens <i>Thanh-Tuan Pham, Thanh Phuong Nguyen</i>	358
Skin Lesion Segmentation based on Integrating EfficientNet and Residual block into U-Net Neural Network <i>Duy Khang Nguyen, Thi-Thao Tran, Cong Phuong Nguyen, Van-Truong Pham</i>	366
An Educational Transformative Sustainability Model Based On Modern Educational Technology <i>Xuan Thanh Pham, Anh Tho Mai, Anh Tuan Ngo</i>	372
A Novel Binning Algorithm Using Topic Modelling and k-mer Frequency on Groups of Non-Overlapping Short Reads <i>Hoang D. Quach, Hoang T. Lam, Dang H. N. Nguyen, Phuong V. D. Van, Van Hoai Tran</i>	380
Study on the Influence of Diaphragm Wall on the Behavior of Pile Raft Foundation <i>Van Qui Lai, Quoc Thien Huynh, Nhat Hoang Vo, Chung Nguyen Van</i>	387
A Theoretical and Numerical Study of Ultrasonic Waves in Laminated Composites for Nondestructive Evaluation of Structures <i>Duy Kien Dao, Duchtho Le, TruongGiang Nguyen, Hoai Nguyen, Duc Chinh Pham, Haidang Phan</i>	392
Guided Wave Propagation in a Layered Half-Space Structure of Anisotropic Materials <i>Duy Kien Dao, Duchtho Le, Quang Hung Le, Minh Tuan Nguyen, Haidang Phan, Duc Chinh Pham</i>	398
City-Scale Electricity Demand Forecasting using a Gaussian Process Model <i>Phong T. T. Nguyen, Lance Manuel</i>	405
An Experimental on Heat Transfer Characteristics of the Cascade Heat Exchanger in Refrigeration System Using R32/CO2 <i>Thanhtrung Dang, Vanloi Nguyen, Hoangtuan Nguyen</i>	413
Research on using PSO Algorithm to Optimize Controlling of Regenerative Braking Force Distribution in Automobile <i>Tung Duong Tuan, Dung Do Van, Thinh Nguyen Truong</i>	418
Design Depth Controller for Hybrid Autonomous Underwater Vehicle using Backstepping Approach <i>Nguyen-Nhut-Thanh Pham, Ngoc-Huy Tran, Thien-Phuong Ton, Thien-Phuc Tran</i>	424
Implementation and Enhancement of Set-Based Guidance by Velocity Obstacle along with LiDAR for Unmanned Surface Vehicles <i>Ngoc-Huy Tran, Minh-Hung Vu, Tu-Cuong Nguyen, Minh-Tam Phan, Quang-Ha Pham</i>	430
Design Integrated Staff Welcoming and Administration System Based on Facial Recognition for Smart University <i>Dat Tan La, Huy Quang Tran, Nhat Tien Le, Quang Luong Nguyen, Thu Thi Anh Nguyen, Tuan Van Pham</i>	436
Saliency prediction for 360-degree video <i>Chuong H. Vo, Jui-Chiu Chiang, Duy H. Le, Thu T.A. Nguyen, Tuan V. Pham</i>	442
Heart Rate Estimation Based on Facial Image Sequence <i>Dao Q. Le, Wen-Nung Lie, Quynh Nguyen Quang Nhu, Thu T.A. Nguyen</i>	449

Stability Analysis of an Islanded Microgrid Using Supercapacitor-based Virtual Synchronous Generator <i>Hong Viet Phuong Nguyen, Van Tan Nguyen, Binh Nam Nguyen, Thi Bich Thanh Truong, Huu Dan Dao, Quoc Cuong Le</i>	454
3D Numerical Simulation Study of a Pre-Heater Used in Solid Oxide Fuel Cell Technology <i>XuanVien Nguyen, TrangDoanh Nguyen, AnQuoc Hoang, MinhHung Doan, ThiNhung Tran</i>	461
Deep Learning Based Semantic Segmentation for Nighttime Image <i>Hien T.T Bui, Duy H. Le, Thu T.A Nguyen, Tuan V. Pham</i>	466
Proposed Smart University Model As A Sustainable Living Lab For University Digital Transformation <i>Tuan V. Pham, Anh Thu T. Nguyen, Thanh Dinh Ngo, Duy H. Le, Khai C.V. Le, Thuong H.N. Nguyen, Huy Q. Le</i>	472
Characteristics of Recycled Reinforced Concrete at High Temperatures <i>Nguyen Thanh Hung, Dao Duy Kien, Nguyen Van Khoa, Tran Minh Hieu, Doan Dinh Thien Vuong</i>	480
The Influence of Raft Thickness on the Behaviour of Piled Raft Foundation <i>Tong Nguyen, Phuong Le, Viet Tran</i>	483
Parallel Multi-Population Technique For Meta-Heuristic Algorithms On Multi Core Processor <i>Nguyen Tien Dat, Cao Van Kien, Ho Pham Huy Anh, Nguyen Ngoc Son</i>	489
Adaptive MIMO Fuzzy Controller for Double Coupled Tank System Optimizing by Jaya Algorithm <i>Cao Van Kien, Nguyen Ngoc Son, Ho Pham Huy Anh</i>	495
A New Approach for Analyzing and Predicting Carbon Dioxide Emissions: Case study of Vietnam <i>Le Thi Giang, Khuu Manh Dat, Nguyen Xuan Hiep, Sam Nguyen-Xuan</i>	500
Proposed Research on Saline-Water Distillation for Living by Utilizing Waste Heat from Industrial Steam Boilers <i>Ngoc Han Pham, Van Tuyen Nguyen</i>	505
Surface Roughness Optimization for Grinding Parameters of SKS3 Steel on Cylindrical Grinding Machine <i>Thi-Minh Pham, Huy-Tuan Pham, Van-Khien Nguyen, Quang-Khoa Dang, Duong Thi Van Anh</i>	511
Study on INTOC Waterproofing Technology for Basement of High-Rise Buildings <i>Sy-Hung Nguyen, Thanh-Tich Do, Julien Ambre</i>	516
Influence of Heating Temperature in Thermal Oxidation to Prepare Titanium Oxide /Aluminum-doped Zinc Oxide Films for Multi-functional-energy-saving Glass <i>Shang-Chou Chang, Tsung-Han Li, Huang-Tian Chan</i>	523
Comparative Studies of Different Methods for Short-term Locational Marginal Price Forecasting <i>Ying-Yi Hong and Rolando Pula</i>	527
Navigation Technology Using Inertial Elements to Compensate for GPS Signal-Shaded in Real Time <i>Guo-Shing Huang, Po-Chun Hsu, Ming-Cheng Kao</i>	533
An Automatic Approach for Estimation of CPR Signal using Thoracic Impedance <i>Van-Truong Pham, Thi-Thao Tran</i>	538
A Study on Patterns of Neural Activity Generation from A Bio-realistic Cerebellum Neural Network <i>Vo Nhu Thanh, Pham Anh Duc, Le Hoai Nam, Dang Phuoc Vinh, Tran Ngoc Hai</i>	544

Integration of MicroSCADA SYS600 9.4 into Distribution Automation System <i>The Khanh Truong, Kim Hung Le, Minh Quan Duong, Tue Truong-Bach, Van Phuong Vo</i>	549
A Study on Urban Traffic Congestion Using Simulation Approach <i>Phan Thi Kim Phung, Nguyen Truong Thi, Vo Thi Kim Cuc</i>	555
Optimizing Warehouse Storage Location Assignment Under Demand Uncertainty <i>Nguyen Truong Thi, Phan Thi Kim Phung, Tran Thi Tham</i>	562
Optimization Design of a Compliant Tension Spring <i>Minh Phung Dang, Hieu Giang Le, Xuan Hoang Vo, Thanh-Phong Dao</i>	569
Application of Fuzzy Control Algorithm to Start a Large -Capacity Synchronous Motor <i>Quoc Hung Duong, Huu Cong Nguyen, The Cuong Nguyen, Hong Quang Nguyen</i>	574
Mobile learning in non-English Speaking Countries: Designing a Smartphone Application of English Mathematical Terminology for Students of Mathematics Teachers Education <i>Bui Anh Tuan, Lam Minh Huy, Nguyen Hieu Thanh, Tieu Ngoc Tuoi, Huynh Tuyet Ngan</i>	582
Fire Resistance Evaluation of Reinforced Concrete Structures <i>Dao Duy Kien, Do Van Trinh, Khong Trong Toan, Le Ba Danh</i>	588
Digital Economy: Overview of Definition and Measurement Criteria <i>Nguyen Thi Thanh Van, Nguyen Thien Duy</i>	593
Receptionist and Security Robot Using Face Recognition with Standardized Data Collecting Method <i>Quang-Minh Ky, Dung-Nhan Huynh, My-Ha Le</i>	597
An efficient evolutionary algorithm for joint optimization of maintenance grouping and routing <i>Ho Si Hung Nguyen, Phuc Do, Hai-Canh Vu, Thanh Bac Le, Kim Anh Nguyen</i>	604
Using Dual-use Electronic Lectures in E-learning: An Empirical Study of Teaching and Learning Mathematics at Vietnamese High Schools <i>Bui Anh Tuan, Tran Thi Thu Thao, Nguyen Ngoc Phuong Anh, Le Thanh Dien</i>	612
Biometric Image Recognition For Secure Authentication Based on FPGA : A survey <i>Huu Q Tran, Van Thai Nguyen</i>	618
A novel two-variable model for bending analysis of laminated composite beams <i>Xuan-Bach Bui, Trung-Kien Nguyen, Quoc-Cuong Le, T. Truong-Phong Nguyen</i>	624
Water, Urban Morphology, Urbanization and Sustainable Development: Case study of the Nhieu Loc – Thi Nghe canal area in Ho Chi Minh city <i>Xuan Son Do</i>	629
An Investigation on Efficiency of Magnetic Assisted Generator <i>Vu-Lan Nguyen, Huann-Ming Chou, Chang-Ren Chen, Bo-Jun Zheng, Kuo-Chen Yang</i>	637
Microstructure and Hardness of Borided Layer on SKD61 Steel <i>Nga Thi-Hong Pham, Long Nhut-Phi Nguyen, The-San Tran</i>	641
An Experimental Study on The Performance of An Air Conditioning System using CO2 Refrigerant with The Actual Power Input of 440W <i>Thanhtrung Dang, Tronghieu Nguyen</i>	645

Redundant Relay Protection Devices for Power Systems Reliability Model <i>Andrey Trofimov, Alexandra Khalyasmaa</i>	651
Forecasting the Solar Power Plants Generation in a Meteorological Data-Constrained Environment <i>Andrey Tokarev, Alexandra Khalyasmaa</i>	657
Vehicle Body Design and Analysis Aerodynamic by Flow Simulation <i>Phu Thuong Luu Nguyen, Van Dung Do</i>	662
A Preliminary Study of a Two Stroke Free- Piston Engine for Electricity Generation <i>Nguyen Huynh Thi, Nguyen Van Trang, Huynh Thanh Cong, Huynh Van Loc, Dao Huu Huy, Ngo Duc Huy</i>	669
Numerical Study on Gas-Assisted Mold Temperature Control with the Application of Air Cover for Improving the Cavity Temperature in Injection Molding Process <i>Son Minh Pham, Minh The Uyen Tran and Thanh Trung Do</i>	673



**2020 5<sup>th</sup> International Conference on Green Technology  
and Sustainable Development (GTSD)**

**November 27-28, 2020 • Ho Chi Minh City, Vietnam**

**COPYRIGHT**

**Part Number: CFP20K53-ART**

**ISBN: 978-1-7281-9982-5**

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyrights Manager at [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org). All rights reserved. Copyright ©2020 by IEEE.

## ORGANIZING COMMITTEE INFORMATION

### **Honorary Chairs:**

Prof. Do Van Dung  
HCMC University of Technology and Education, Vietnam

Prof. Nguyen Ngoc Thanh  
Wroclaw University of Science and Technology, Poland

Prof. Nguyen Ngoc Vu  
The University of Da Nang, Vietnam

### **General Chairs**

Assoc.Prof. Le Hieu Giang  
HCMC University of Technology and Education, Vietnam

Assoc.Prof. Ngo Van Thuyen  
HCMC University of Technology and Education, Vietnam

Assoc.Prof. Le Quang Son  
The University of Da Nang, Vietnam

### **Organizing Chairs**

Assoc.Prof. Hoang An Quoc  
HCMC University of Technology and Education, Vietnam

Assoc.Prof. Nguyen Le Hung  
The University of Da Nang, Vietnam

### **Program Chairs**

Prof. Wen-June Wang  
National Central University, Taiwan

Prof. Yo-Ping Huang  
National Taipei University of Technology, Taiwan

### **Publication Chair**

Assoc.Prof. Do Thanh Trung  
HCMC University of Technology and Education, Vietnam

### **Secretaries**

Chau Ngoc Thin  
HCMC University of Technology and Education, Vietnam

Vu Thi Thanh Thao  
HCMC University of Technology and Education, Vietnam

## AUTHORS INDEX

- A**
- Alang Tho, 286  
Alberto Berizzi, 41  
Alessandro Bosio, 41  
Alexander Kies, 253  
Alexandra Khalyasmaa, 651, 657  
Andrea Morotti, 41  
Andrea Vicario, 41  
Andrey Tokarev, 657  
Andrey Trofimov, 651  
Anh Phan Tuan, 66  
Anh Thang Le, 192  
Anh Thi-Kim Tran, 214  
Anh Tho Mai, 372  
Anh Thu T. Nguyen, 472  
Anh Tuan Ngo, 372  
Anh-Thang Le, 175, 209  
AnQuoc Hoang, 461
- B**
- Bach-Duong Pham, 321  
Bartolomeo Greco, 41  
Ba-Viet Ngo, 219, 344  
Binh Nam Nguyen, 454  
Blerand Greicevci, 88  
Bo-Jun Zheng, 637  
Bruno Schyska, 253  
Bui Anh Tuan, 582, 612  
Bui Ha Duc, 180  
Bui Manh Huy, 241  
Bui Pham Duc Tuong, 1
- C**
- Cao Van Kien, 489, 495  
Chang-Ren Chen, 637  
Chao-Lung Hwang, 264, 268  
Chung Nguyen Van, 387  
Chuong H. Vo, 442  
Cong Binh Phan, 124  
Cong Phuong Nguyen, 366
- D**
- Dang H. N. Nguyen, 380  
Dang My Duyen Nguyen, 247  
Dang Phuoc Vinh, 544  
Dang Thi Truc Linh, 232  
Dang Xuan Ba, 241  
Dang-Khoa Tran, 219  
Dao Duy Kien, 317, 480, 588  
Dao Huu Huy, 669  
Dao Q. Le, 449  
Dat Tan La, 436  
Detlev Heinemann, 253  
Dinh Cao Tri, 326  
Dinh Son Nguyen, 157  
Dinh Thanh Viet, 253  
Dinh-Duong Le, 41  
Dinh-Nhon Truong, 296  
Do Van Trinh, 588  
Doan Dinh Thien Vuong, 480  
Duc Chinh Pham, 392, 398  
Duc Nguyen Minh, 162  
Duc Thien Tran, 185  
Duc Tri Do, 53  
Duc-Dung Vo, 137, 344  
Ductho Le, 392, 398  
Duc-Tri Do, 78  
Dung Do Van, 418  
Dung Ong Mau, 66  
Dung-Nhan Huynh, 597  
Duong Minh Quan, 253  
Duong Thi Van Anh, 511  
Duy H. Le, 442, 466, 472  
Duy Khang Nguyen, 366  
Duy Kien Dao, 392, 398  
Duy Liem Nguyen, 98  
Duy-Hai Vo, 264, 298  
Duy-Khanh Ha, 280  
Duy-Liem Nguyen, 30, 36, 302, 312
- F**
- Fabian Hofmann, 253  
Francois Christophe, 88
- G**
- Gaetano Iannarelli, 41  
Guo-Shing Huang, 533
- H**
- H.T. Tai Nguyen, 302, 312  
Ha Tran Thu, 133  
Hai-Canh Vu, 604  
Haidang Phan, 392, 398  
Hannu Fagerlund, 88  
Hien T.T Bui, 466  
Hieu Giang Le, 78, 569  
Ho Pham Huy Anh, 489, 495  
Ho Si Hung Nguyen, 604  
Hoai Nam Le, 157  
Hoai Nguyen, 392  
Hoang D. Quach, 380  
Hoang T. Lam, 380  
Hoang Thi Ngoc Mai, 340  
Hoang-Anh Nguyen, 264, 268  
Hoangtuan Nguyen, 413  
Hong Nga Thi Pham, 5  
Hong Quang Nguyen, 574  
Hong Viet Phuong Nguyen, 454  
Hong-Xuyen Ho Thi, 259  
Huang-Tian Chan, 523  
Huann-Ming Chou, 637  
Hung Ngoc Nguyen, 203  
Hung Nguyen, 272

Huu Cong Nguyen, 574

Huu Dan Dao, 454

Huu Q Tran, 618

Huu Vinh Nguyen, 272

Huy Q. Le, 472

Huy Quang Tran, 436

Huynh Thanh Cong, 669

Huynh Tuyet Ngan, 582

Huynh Van Loc, 669

Huynh Vinh Nghi, 241

Huynh-Tan-Tai Nguyen, 30

Huy-Tuan Pham, 511

Hyug-Moon Kwon, 236

## J

Jarmo Havula, 88

Jin-Gyun Kim, 59

John Burgess, 286

Jui-Chiu Chiang, 442

Julien Ambre, 516

## K

Khac Hieu Nguyen, 94

Khai C.V. Le, 472

Khanh Son Trinh, 119

Khanh-Dung Tran Thi, 264, 268

Khoe Dang Dinh, 203

Khong Trong Toan, 588

Khuu Manh Dat, 500

Kim Anh Nguyen, 604

Kim Hung Le, 272, 549

Kuo-Chen Yang, 637

## L

Lam Chuong Vo, 334

Lam Minh Huy, 582

Lan Anh Nguyen Thi, 259

Lance Manuel, 405

Le Ba Danh, 588

Le Duc Hanh, 9

Le Hoai Nam, 544

Le Hoai Phuong, 9

Le Minh Nhut, 232

Le Thanh Dien, 612

Le Thanh Phuc, 326

Le Thi Giang, 500

Le Viet Thinh, 168

Linh T. K. Vu, 203

Loc-Phat Truong, 321

Loi Dang-Huu, 59

Long Nhut-Phi Nguyen, 641

Luan Vu Truong Nguyen, 334

Lueder von Bremen, 253

## M

Ma Phuoc Khanh, 253

Manh Cuong Le, 349

Manh Son Tran, 185

Man-Yin Wu, 128

Markus Schlott, 253

Milan Gautam, 88

Ming-Cheng Kao, 533

Minh Hao Hoang, 291

Minh Le Hoang, 133

Minh Phung Dang, 569

Minh Quan Duong, 549

Minh The Uyen Tran, 673

Minh Tien Cao, 272

Minh Tuan Nguyen, 398

Minh Vuong Le, 272

Minh-Duc Nguyen, 48, 175

MinhHung Doan, 461

Minh-Hung Vu, 430

Minh-Khai Nguyen, 78

Minh-Tam Phan, 430

Minh-Thien Duong, 349, 353

Min-Kyoung Kim, 30

Mitiku Damtie Yehualaw, 264

Mitiku Damtie Yehualaw, 268

Moonyong Lee, 334

My-Ha Le, 349, 353, 597

## N

Nga Thi Vo, 291

Nga Thi-Hong Pham, 641

Nga Vu Thi Hong, 66

Ngo Duc Huy, 669

Ngoc Han Pham, 505

Ngoc Thanh Tran, 98

Ngoc Thien Tran, 5

Ngoc-Duong Nguyen, 73

Ngoc-Huy Tran, 424, 430

Nguyen Duc Anh, 103

Nguyen Duc Tuyen, 168

Nguyen Hieu Thanh, 582

Nguyen Hong Ngoc Linh, 340

Nguyen Huynh Thi, 669

Nguyen Khac Hieu, 114

Nguyen Manh Tuan, 192

Nguyen Ngoc Phuong Anh, 612

Nguyen Ngoc Son, 489, 495

Nguyen Nhat Huy, 340

Nguyen Thai Anh, 340

Nguyen Thanh Hung, 317, 480

Nguyen Thanh Son, 19

Nguyen Thi Anh Van, 114, 198

Nguyen Thi Hoang Anh, 198

Nguyen Thi Thanh Van, 593

Nguyen Thien Duy, 593

Nguyen Tien Dat, 489

Nguyen Truong Thi, 555, 562

Nguyen Van Hiep, 185

Nguyen Van Khoa, 317, 480

Nguyen Van Trang, 669

Nguyen Vu Quynh, 24

Nguyen Xuan Hiep, 500

Nguyen-Nhut-Thanh Pham, 424

Nguyen-Thao Thach, 175

Nhat Hoang Vo, 387

Nhat Tien Le, 436

Nhu Thanh Vo, 157

Nhung Thi-Tuyet Hoang, 214

## P

Pei-Jun Lee, 296  
Pham Anh Duc, 544  
Pham Duc Thien, 317  
Pham Ngoc Son, 103  
Pham Tan Phat, 241  
Phan Duc Huynh, 1  
Phan Thanh Nhan, 15  
Phan Thi Kim Phung, 555, 562  
Phan Thi Thu Thuy, 214  
Phong T. T. Nguyen, 405  
Phu Thuong Luu Nguyen, 662  
Phuc Do, 604  
Phuong Le, 483  
Phuong V. D. Van, 380  
Phuong-Doanh Huynh, 209  
Po-Chun Hsu, 533

## Q

Quang Hung Le, 398  
Quang Luong Nguyen, 436  
Quang-Ha Pham, 430  
Quang-Huy Vu, 321  
Quang-Khoa Dang, 511  
Quang-Minh Ky, 597  
Quoc Cuong Le, 454  
Quoc Hung Duong, 574  
Quoc Khanh Tran, 98  
Quoc Thien Huynh, 387  
Quoc-Cuong Le, 624  
Quyen Le Ly Quyen, 66  
Quynh Nguyen Quang Nhu, 449

## R

Rolando Pula, 527

## S

Sam Nguyen-Xuan, 500  
Shang-Chou Chang, 523  
Sheng-Huei Lee, 128  
Son Minh Pham, 673

Son Tran Minh, 133  
Soo-Yong Kim, 280  
Stéphane Bouissou, 224  
Sy-Hung Nguyen, 224, 516

## T

T. Truong-Phong Nguyen, 624  
Tai Tran Thanh, 236  
Tan Hung Nguyen, 272  
Thai Nguyen Van, 133  
Thanh Bac Le, 604  
Thanh Chau-Dinh, 59  
Thanh Dinh Ngo, 472  
Thanh Hai Quach, 53  
Thanh Hai Tuan Tran, 157  
Thanh Phuong Nguyen, 358  
Thanh Trung Do, 673  
Thanh Tung Pham, 247  
Thanh-Hai Nguyen, 137, 219, 344  
Thanh-Hai Quach, 78  
Thanh-Hung Nguyen, 209  
Thanh-Kiet Phan, 175  
Thanh-Nghia Nguyen, 137  
Thanh-Nhan Phan, 302  
Thanh-Phong Dao, 569  
Thanh-Tam Nguyen, 344  
Thanh-Thanh Ngo-Quang, 153  
Thanh-Tich Do, 516  
Thanhtrung Dang, 413, 645  
Thanh-Tuan Pham, 358  
The Cuong Nguyen, 574  
The Khanh Truong, 549  
The-San Tran, 641  
Thi Anh Van Nguyen, 94  
Thi Bich Dung Phung, 109  
Thi Bich Thanh Truong, 454  
Thi Duy Hanh Le, 119  
Thi Hao Cao, 291  
Thien-An Nguyen, 175  
Thien-Phuc Tran, 424  
Thien-Phuong Ton, 424

Thi-Minh Pham, 511  
Thi-My Ngo, 264  
Thi-Ngoc-Han Vuong, 36, 78  
Thi-Ngoc-Hieu Phu, 148  
Thinh Nguyen Truong, 418  
Thinh Vinh Le, 141  
ThiNhung Tran, 461  
Thi-Phuong-Huyen Tran, 224  
Thi-Thao Tran, 366, 538  
Thu T.A. Nguyen, 442, 466, 449  
Thu Thi Anh Nguyen, 436  
Thuong H.N. Nguyen, 472  
Tien Hoang Nguyen, 272  
Tien-Tho Do, 36, 302, 312  
Tieu Ngoc Tuoi, 582  
Ton Duc Do, 296  
Tong Nguyen, 483  
Tran Hoai Nam, 241  
Tran Minh Hieu, 480  
Tran Ngoc Hai, 544  
Tran Nhat Huy, 180  
Tran Thi Tham, 562  
Tran Thi Thu Thao, 612  
Tran Thien Huan, 141  
Tran Thuy Ai Phuong, 198  
Tran Tien Khoi, 340  
TrangDoanh Nguyen, 461  
Tri Thuong Ngo, 98  
Tri Tran Quang, 286  
Tronghieu Nguyen, 645  
Trung-Hieu Nguyen, 148, 153  
Trung-Kien Nguyen, 73, 624  
Truong-Duy Nguyen, 137, 219  
TruongGiang Nguyen, 392  
Tsung-Han Li, 523  
Tu Nguyen Thanh, 162, 236  
Tu Tran, 286  
Tuan V. Pham, 442, 466, 472  
Tuan Van Pham, 436  
Tuan-Anh Nguyen, 109  
Tu-Cuong Nguyen, 430

Tue Truong-Bach, 549  
Tung Duong Tuan, 418  
Tuyen Dinh Quang, 66

## V

Van Dung Do, 662  
Van Hien Bui, 53  
Van Hoai Tran, 380  
Van Phuong Vo, 549  
Van Qui Lai, 387  
Van Tan Nguyen, 454  
Van Thai Nguyen, 618  
Van Tuyen Nguyen, 505  
Van-Binh Nguyen, 353  
Van-Dung Do, 153  
Van-Khien Nguyen, 511  
Van-Khoa Nguyen, 280  
Van-Khoa Pham, 82

Vanloi Nguyen, 413  
Van-Ngoc Tong, 296  
Van-Phong Vu, 296  
Van-Thuyen Ngo, 153, 219, 296  
Van-Tin Luu, 148  
Van-Truong Pham, 366, 538  
Viet Anh Truong, 53  
Viet Tran, 483  
Viet-Anh Truong, 78  
Viet-Cuong Huynh, 148  
Vinh-Thanh Tran, 78  
Vo Duy Cong, 9  
Vo Nhu Thanh, 544  
Vo Thi Kim Cuc, 555  
Vo Van Phuong, 253  
Vu Xuan Son Huu, 168  
Vu-An Tran, 268  
Vu-Hoang Tran, 148, 153

Vu-Lan Nguyen, 637  
Vu-Tu Tran, 30, 36, 302, 312

## W

Wen-June Wang, 296  
Wen-Nung Lie, 449

## X

Xuan Hoang Vo, 569  
Xuan Son Do, 629  
Xuan Thanh Pham, 372  
Xuan Truong Luong, 53  
Xuan-Bach Bui, 624  
XuanVien Nguyen, 461

## Y

Ying-Yi Hong, 128, 527

# Numerical Study on Vibration Control of Structures Using Multi Tuned Liquid Dampers with High Mass Ratio

Bui Pham Duc Tuong  
Faculty of Civil Engineering  
HCM University of Technology and Education  
Ho Chi Minh city, Viet Nam  
tuongbpd@hcmute.edu.vn

Phan Duc Huynh  
Faculty of Civil Engineering  
HCM University of Technology and Education  
Ho Chi Minh city, Viet Nam  
corresponding author:  
huynhpd@hcmute.edu.vn

**Abstract**— Multiple tuned liquid dampers (MTLDs) with mass ratio 5% presented in this study. The results of numerical simulation based on the equivalent mechanical model are carried out on the multi-degree-of-freedom (MDOF) affected by harmonic forces. A parametric study is performed in the frequency domain to investigate the dynamic characteristics and effectiveness of MTLDs. The influence of mass ratio, number of TLDs, normalized frequency band width, and number of degree-of-freedom (DOF) are surveyed. The results show that the control method by using MTLDs in case of high mass ratio is very effective.

**Keywords**—Multi tuned liquid damper, Structural control, Mass ratio, Mechanical model.

## I. INTRODUCTION

TLD has been introduced as an energy-consuming device to reduce structural vibrations. TLDs, which include a liquid storage tank, are a passive control device to re-duce structural vibrations when external forces exert such as wind and earthquakes. The basic sloshing frequency of a liquid is usually adjusted to the fundamental frequency of the main structure. When the TLD is stimulated by the movements of the main structure, the fluid inside the tank begins to slosh, transmitting inertial forces to the structure, thus absorbing and dissipating energy. The energy consumption in TLDs occurs through the phenomenon of breaking waves, the geometry of TLDs, the viscosity of liquids ...

TLD can be classified into two types: shallow water and deep water. This classification is based on the ratio of the depth of the liquid to the length of the tank in the direction of propagation. The idea of using a TLD to reduce vibration in civil engineering structures has been researched and developed since the mid-1980s. [1] is the first to suggest using a rectangular container filled with two immiscible liquids to reduce oscillation through the movement of the interface. TLD has been shown to effectively control the response of structures caused by wind [2] [3], and have successfully implemented applications in many high-rise structures [4]. In addition, the TLD system is proposed to control earthquake excitation [5]. Experimental and theoretical research [6][7][8] show that TLDs are effective in controlling the response of structures under the effect of earthquakes. In addition, TLDs are also proposed to be used on cable stayed bridges [9] and wind turbines [10]. [11] studied some kinematic parameters of TLD. Parameters include sloshing frequency, fluid viscosity, bottom roughness and mass ratio between TLD and structure.

The effects of flexible wall and nonlinear parameters of TLD control results were also studied [12][13][14][15].

Although the application of a single TLD may be useful in reducing structural responses under external excitation, it comes with a number of limitations such as the ability of the TLD to respond when the frequency changes, damping ratio, and uncertainty about the dynamics of the main structure. To overcome the limitations and improve the performance of a TLD, a multiple TLD system (MTLDs) is proposed.

MTLDs include many TLDs with similar or different dynamics properties. MTLDs can be located on one floor or distributed on different floors. The sloshing frequency of each TLD can be easily changed by changing the fluid depth.

The MTLDs system with mass ratio 5% is presented in this study. The numerical simulation results are based on an equivalent mechanical model performed on the model affected by harmonic forces. The influence of mass ratio, number of TLDs, normalized frequency band width, and number of floors are surveyed. The results show that the control model of MTLDs is very effective.

## II. EQUIVALENT MECHANICAL MODEL OF TLD

Consider a hard-walled TLD –  $i$ . The height of fluid when the TLD is not moving is  $H_i$ . The width of the fluid is  $2l_i$ . When TLD is under the effect of horizontal acceleration,  $\ddot{X}_0(t)$ , the fluid mass will also move horizontally. The equivalent mass-damper-stiffness system of TLD is shown in Fig. 1b. Mass values are determined by the equation (1):

$$m_i = M_i^{TLD} \frac{2\left(\frac{H_i}{l_i}\right)^2 \tanh(\beta_i H_i)}{(\beta_i H_i)^3} \quad (1)$$

$$m_{0i} = M_i^{TLD} \left(1 - \frac{m_i}{M_i^{TLD}}\right) \quad (2)$$

With  $\beta_i = \frac{\pi}{2l_i}$ , and  $M_i^{TLD}$  is the mass of fluid.

$$M_i^{TLD} = 2\rho l_i H_i B_i \quad (3)$$

$\rho$  is the density of the fluid;  $B_i$  is the thickness of the fluid. The equivalent stiffness of TLD is found by the formula:

$$k_i = m_i \omega_i^2 \quad (4)$$

The equivalent damping of the TLD is determined by the damping coefficient  $\zeta_i$  of the fluid contained in the TLD,

$$c_i = 2m_i \omega_i \zeta_i \quad (5)$$

In which the damping ratio is determined based on the formula

$$\zeta_i = \left( \frac{1}{H_i} + \frac{1}{B_i} \right) \sqrt{\frac{\vartheta}{2\omega_i}} \quad (6)$$

$\vartheta$  is the kinematic viscosity coefficient of fluid.

The horizontal movement frequency  $\omega_i$  of fluid is determined by the formula

$$\omega_i = \sqrt{(g\beta_i \tanh(\beta_i H_i))} \quad (7)$$

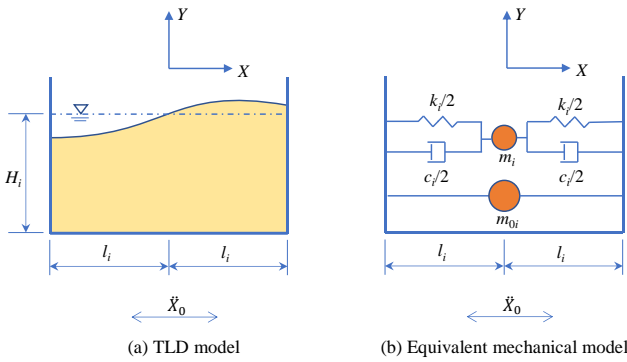


Fig. 1. TLD in horizontal motion

### III. THE MOTION EQUATION OF THE STRUCTURE MDOF AND MTLD

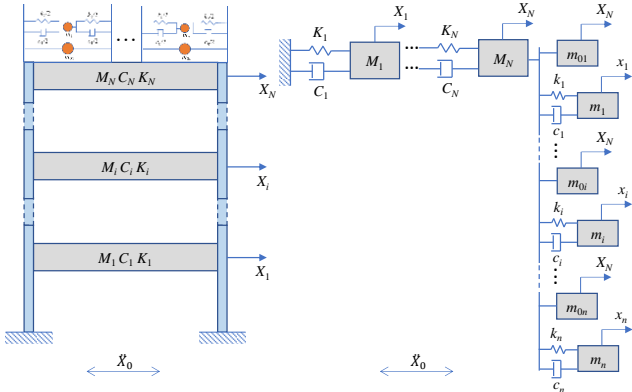


Fig. 2. System of multiple TLDs for high-rise structures

The main structure is a multi-story building modeled into a mass-stiffness-damping system. Structural properties such as mass, stiffness, and damping of each floor are identical. The model of a N-floor building with n-TLDs located on the top floor of the main structure is shown in Fig. 2. The frequency of the n-TLD is distributed around the natural frequency of the main structure. The main structure is symmetrical and has uniform mass  $M_1 = M_i = M_N = M_0$ ; horizontal stiffness  $K_1 = K_i = K_N = K_0$ ; horizontal damping  $C_1 = C_i = C_N = C_0$ .

The motion equation of the MDOFs multilevel structure associated with many MTLD control systems is presented in the following general form:

$$\mathbf{M}\ddot{\mathbf{X}} + \mathbf{C}\dot{\mathbf{X}} + \mathbf{K}\mathbf{X} = \mathbf{P} \quad (8)$$

In equation (8),  $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_N \ x_1 \ x_2 \ \dots \ x_n]$  is a horizontal displacement vector with  $X_i$  = displacement of the  $i^{th}$  floor;  $\mathbf{P}$  is the excitation vector caused by the earthquake. The mass, damping, and stiffness matrices in equation (8) are defined as follows.

$$\mathbf{M} = M_0 \bar{\mathbf{M}} \quad (9)$$

$$\bar{\mathbf{M}} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & (1 + \sum_{i=1}^n r_{0i}^m) & \\ & & & r_1^m & \ddots \\ & & & & & r_n^m \end{bmatrix} \quad (10)$$

$$\mathbf{C} = C_0 \bar{\mathbf{C}} \quad (11)$$

$$\bar{\mathbf{C}} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & & & \\ & & \ddots & & \\ & & & (1 + \sum_{i=1}^n r_i^c) & -r_1^c & \dots & -r_n^c \\ & & & -r_1^c & r_1^c & & \\ & & & \vdots & & \ddots & \\ & & & -r_n^c & & & r_n^c \end{bmatrix} \quad (12)$$

$$\mathbf{K} = K_0 \bar{\mathbf{K}} \quad (13)$$

$$\bar{\mathbf{K}} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & & & \\ & & \ddots & & \\ & & & (1 + \sum_{i=1}^n r_i^k) & -r_1^k & \dots & -r_n^k \\ & & & -r_1^k & r_1^k & & \\ & & & \vdots & & \ddots & \\ & & & -r_n^k & & & r_n^k \end{bmatrix} \quad (14)$$

With  $\bar{\mathbf{M}}$ ,  $\bar{\mathbf{C}}$ , and  $\bar{\mathbf{K}}$  normalized matrices of mass, damping, and stiffness, respectively;

$$r_{0i}^m = \frac{m_{0i}}{M_0} = r_i^{m,TLD} \left( 1 - \frac{16}{\pi^3} \frac{\tanh(\frac{\pi}{2}\eta_i)}{\eta_i} \right) \quad (15)$$

$$r_i^m = \frac{m_i}{M_0} = r_i^{m,TLD} - r_{0i}^m \quad (16)$$

$$r_i^k = \frac{k_i}{K_0} = r_i^m \gamma_i \quad (17)$$

With  $\eta_i = H_i/l_i$ ;  $r_i^{m,TLD} = M_i^{TLD}/M_0$ ;  $\gamma_i = \omega_i/\omega_0$  is the ratio of the frequency of the TLD-i to the parameter  $\omega_0 = \sqrt{K_0/M_0}$ .

Vector of force due to earthquake acceleration  $\ddot{X}_0(t) = \ddot{X}_{0a}\ddot{X}_0(t)$  is determined by



$$\mathbf{P} = -M_0 \bar{\mathbf{D}}_M \ddot{\mathbf{X}}_{0a} \bar{\mathbf{X}}_0 \quad (18)$$

In equation (18)  $\bar{\mathbf{D}}_M$  is the diagonal vector of matrix  $\bar{\mathbf{M}}$ ,

$$\bar{\mathbf{D}}_M = [1 \quad 1 \quad \cdots \quad (1 + \sum_{i=1}^n r_{0i}^m) \quad r_1^m \quad \cdots \quad r_n^m]^T \quad (19)$$

For a reference response  $X_{0,st}$  or static amplitude defined by

$$X_{0,st} = \frac{M_0 \ddot{X}_{0a}}{K_0} \quad (20)$$

Equation (8) is expressed as a dimensionless form or normalized equations of motion as:

$$\bar{\mathbf{M}} \ddot{\bar{\mathbf{X}}} + 2\zeta_0 \omega_0 \bar{\mathbf{C}} \dot{\bar{\mathbf{X}}} + \omega_0^2 \bar{\mathbf{K}} \bar{\mathbf{X}} = -\omega_0^2 \bar{\mathbf{D}}_M \ddot{\bar{\mathbf{X}}}_0 \quad (21)$$

Where normalized response value  $\bar{\mathbf{X}} = \mathbf{X}/X_{0,st}$ , and  $\zeta_0 = C_0/(2M_0\omega_0)$ .

The equation (21) is solved by Runge-Kutta method to determine the normalized response value  $\bar{\mathbf{X}}$ .

#### IV. NUMERICAL STUDY ON THE INFLUENCE OF PARAMETERS

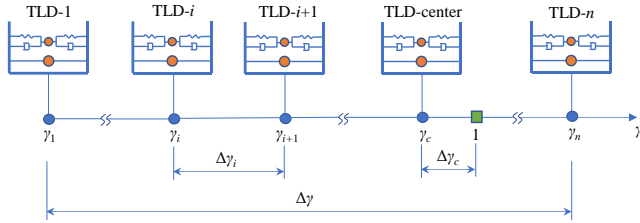


Fig. 3. TLDs distribution and normalized frequency  $\gamma$

As depicted in Fig. 3, the distribution of the normalized frequency  $\gamma_i = \omega_i/\omega_0$  of MTLDs has the following influencing factors: central normalized frequency  $\gamma_c$ , normalized frequency band width  $\Delta\gamma = \gamma_n - \gamma_1$ , the difference between two consecutive normalized frequencies  $\Delta\gamma_i = \gamma_{i+1} - \gamma_i$ , and the deviation  $\Delta\gamma_c$  between the central normalized frequency and 1. In this study,  $\Delta\gamma_c = 0$ .

To investigate the effect of parameters on control results, equation (21) is solved in frequency response with ground motion in the following harmonic form:

$$\ddot{\bar{\mathbf{X}}}_0 = e^{i\omega t} \quad (22)$$

Then the response of equation (21) is written in the form

$$\bar{\mathbf{X}} = \Theta e^{i\omega t} \quad (23)$$

Derivative and substitute to equation (21), the frequency response amplitude is as follows:

$$\Theta\left(\frac{\omega}{\omega_0}\right) = \left[ -\left(\frac{\omega}{\omega_0}\right)^2 \bar{\mathbf{M}} + 2\zeta_0 \left(\frac{\omega}{\omega_0}\right) i \bar{\mathbf{C}} + \bar{\mathbf{K}} \right]^{-1} \bar{\mathbf{D}}_M \quad (24)$$

In this study the damping ratio is  $\zeta_0 = 0.5\%$ . The simulations were carried out with the aim of studying the effects of several parameters of the MTLDs on its performance, i.e. mass ratio, number of TLDs  $n$ , the

normalized frequency band width  $\Delta\gamma$ , and the floor numbers  $N$ .

##### A. Effect of mass ratio

The influence of mass ratio on the control ability of the structure investigated in the case of  $\Delta\gamma = 0.2, n = 1, N = 1$ . The result of amplitude response  $\Theta$  to the change of frequency ratio  $\omega/\omega_0 = [0.8 \div 1.2]$  is shown in Fig. 4. The absence of TLDs is also depicted in the figure. Simulation results show the effectiveness of control using TLDs. With mass ratio increased, control efficiency also increased. But the effect does not change much when the mass ratio is greater than 0.05. In this study, mass ratio = 0.05 model was surveyed to check the performance of the control system.

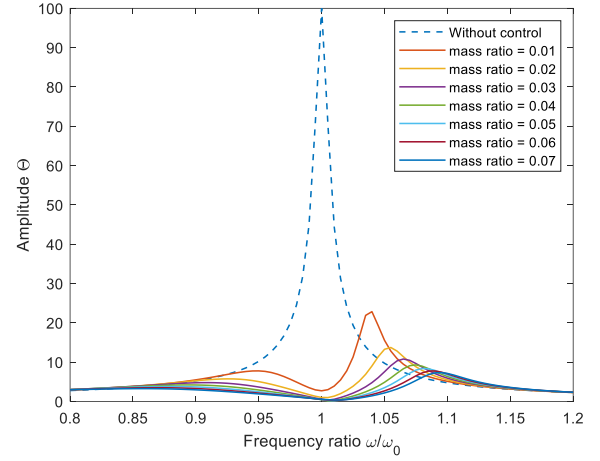


Fig. 4. Effect of mass ratio

##### B. Effect of number of TLDs $n$

Fig. 5 shows numerical simulation results for mass ratio = 0.05. For cases where the number of TLDs is 5, the curve has some local vertices. As the number of TLDs increased to 11, the upper peaks became flatter. When  $n$  increases to 21 or 31, the curve does not change much. It shows that the control efficiency of MTLDs does not change if the number of TLD exceeds a certain value.

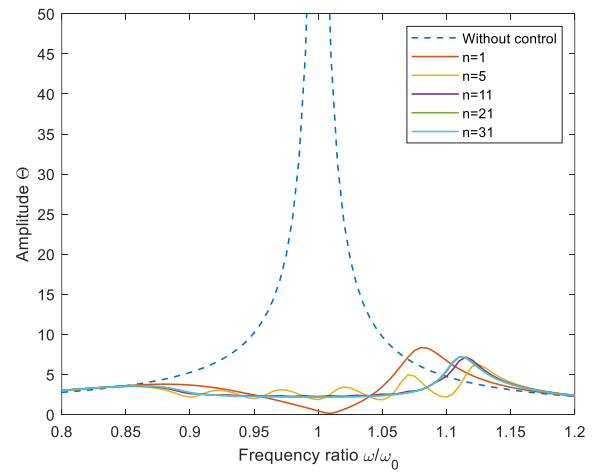


Fig. 5. Effect of number of TLDs in case of mass ratio = 0.05

##### C. Effect of normalized frequency band width $\Delta\gamma$

Simulation results obtained when mass ratio = 0.05. The result is shown in Fig. 6. In case  $\Delta\gamma = 0.3$ , the control value

is not good. Through the above results, we see that there will exist an optimal value of  $\Delta\gamma$  for each case of mass ratio, parameters of the structure such as natural frequency, damping ratio .... The results show that the optimal value of  $\Delta\gamma$  needs more research.

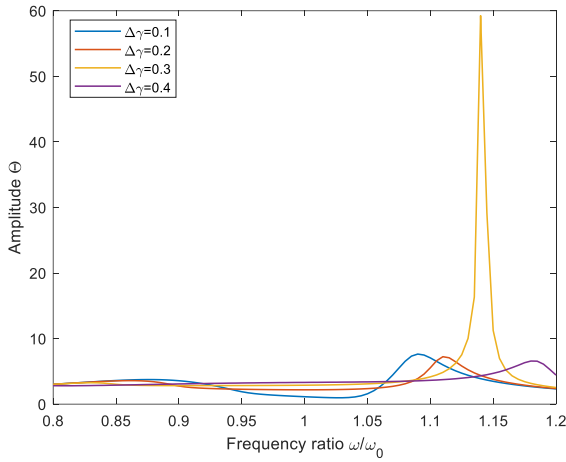


Fig. 6. Effect of normalized frequency band width in case of  $n = 31$

#### D. Effect of TLDs on the number of DOFs

Fig. 7 shows the control effect of TLDs on MDOFs structures. The results show that the number of TLDs will no longer have a large effect when  $n$  is greater than 5. However, with random loads,  $n > 5$  may be effective over a given frequency range.

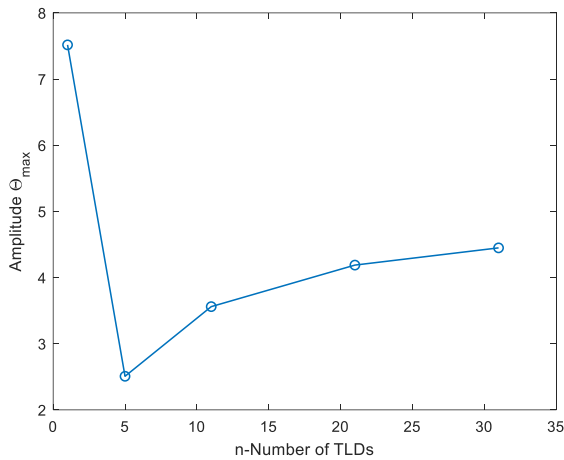


Fig. 7. Effect of control results when  $N = 50$

#### V. CONCLUSIONS

TLDs are effective vibration absorbers when they are applied in practice due to their ease of installation, minimal maintenance and cost effectiveness. In this paper, the use of multiple TLDs in the vibration control of MDOFs structures has been investigated based on an equivalent mechanical model. The influences of the number of TLDs, normalized frequency band width, and degrees of freedom on a control

system with a mass ratio of 0.05 were studied based on numerical methods. The results of this study confirm that multiple TLDs adjusted to mechanical properties of MDOFs system will lead to improved vibration control compared to cases a single TLD used.

#### REFERENCES

- [1] Bauer HF, "Oscillations of immiscible liquids in a rectangular container: A new damper for excited structures," *J Sound Vib*, Vol. 93, pp. 117–133, 1984.
- [2] Fujii K, Tamura Y, Sato T, Wakahara T, "Wind-induced vibration of tower and practical applications of tuned sloshing damper," *J Wind Eng Ind Aerodyn*, Vol. 33, pp. 263–272, 1990.
- [3] C.C. Chang, M. Gu, "Suppression of vortex-excited vibration of tall buildings using tuned liquid dampers," *J Wind Eng Ind Aerodyn*, Vol. 83, pp. 225–237, 1999.
- [4] Ueda T, Nakagaki R, Koshida K, "Suppression of wind-induced vibration by dynamic dampers in tower-like structures," *J Wind Eng Ind Aerodyn*, Vol. 43, pp. 1907–1918, 1992.
- [5] Modi VJ, Welt F, "Damping of wind induced oscillations through liquid sloshing," *J Wind Eng Ind Aerodyn*, Vol. 30, pp. 85–94, 1988.
- [6] Banerji P, Murudi M, Shah A, Popplewell N, "Tuned Liquid Dampers for Control of Earthquake Response," *13th World Conf Earthq Eng*, pp. 587–602, 2000.
- [7] Lee SK, Park EC, Min KW, et al, "Real-time hybrid shaking table testing method for the performance evaluation of a tuned liquid damper controlling seismic response of building structures," *J Sound Vib*, Vol. 302, pp. 596–612, 2007.
- [8] Jin Q, Li X, Sun N, et al, "Experimental and numerical study on tuned liquid dampers for controlling earthquake response of jacket offshore platform," *Mar Struct*, Vol. 20, pp. 238–254, 2007.
- [9] Shum KM, Xu YL, Guo WH, "Wind-induced vibration control of long span cable-stayed bridges using multiple pressurized tuned liquid column dampers," *J Wind Eng Ind Aerodyn*, Vol. 96, pp. 166–192, 2008.
- [10] Zhang Z, Staino A, Basu B, Nielsen SRK, "Performance evaluation of full-scale tuned liquid dampers (TLDs) for vibration control of large wind turbines using real-time hybrid testing," *Eng Struct*, Vol. 126, pp. 417–431, 2016.
- [11] Fujino Y, Pacheco BM, Chaiseri P, Sun LM, "Parametric studies on tuned liquid damper (TLD) using circular containers by free-oscillation experiments," *Doboku Gakkai Rombun-Hokokushu/Proceedings Japan Soc Civ Eng*, pp. 177–187, 1988.
- [12] Tuong BPD, Huynh PD, Bui T-T, Sarhosis V, "Numerical Analysis of the Dynamic Responses of Multistory Structures Equipped with Tuned Liquid Dampers Considering Fluid-Structure Interactions," *Open Constr Build Technol J*, Vol. 13, pp. 289–300, 2019.
- [13] Dziedzic K, Ghosh A, Iwaniec J, et al, "Analysis of tuned liquid column damper nonlinearities," *Eng Struct*, Vol. 171, pp. 1027–1033, 2018.
- [14] Farid M, Gendelman O V., "Response regimes in equivalent mechanical model of moderately nonlinear liquid sloshing," *Nonlinear Dyn*, Vol. 92, pp. 1517–1538, 2018.
- [15] Strand IM, Faltinsen OM, "Linear sloshing in a 2D rectangular tank with a flexible sidewall," *J Fluids Struct*, Vol. 73, pp. 70–81, 2017.
- [16] Li Y, Wang J, "A supplementary, exact solution of an equivalent mechanical model for a sloshing fluid in a rectangular tank," *J Fluids Struct*, Vol. 31, pp. 147–151, 2012.

# Modelling the Distribution of TiC Reinforced Co-Based on Skd61 Steel Surface Prepared by Laser Cladding

Ngoc Thien Tran

Department of Welding and Metal Technology, FME  
HCMC University of Technology and Education  
Ho Chi Minh City, Viet Nam  
thientn@hcmute.edu.vn

Hong Nga Thi Pham

Department of Welding and Metal Technology, FME  
HCMC University of Technology and Education, Viet Nam  
Ho Chi Minh City, Viet Nam  
hongnga@hcmute.edu.vn

**Abstract**— Applying the Laser Cladding method to create composites cladding Co+(10; 20; 30)% TiC on the surface of Hot Die, SKD61 steel. Using Micro-hardness tester and Microscope (SEM) to measure the hardness and analyze the distribution of TiC particles in the clad. The results indicated that the average hardness of the TiC/Co-based composites cladding cross-section Co-based+10%TiC, Co-based+20%TiC, Co-based+30%TiC in order is 589 HV0.2, 788 HV0.2, and 934 HV0.2. All of these claddings have the hardness higher than SKD61 steel in order is 2.6, 3.5, and 4.2 times. From that show that, the hardness of the claddings has a close relationship with percentage of TiC particles. Besides, the regression models exactly show that the hardness of the claddings is:  $y_1 = 23.8438 - 7.3494x$ ,  $y_2 = 9.7389x + 44.6927$ ,  $y_3 = 13.6597x + 35.511$  by 10, 20, 30%wt TiC particles in the claddings, which create the scientific theory to apply from some other carbides to reinforced for the cladding.

**Keywords**— Laser Cladding, TiC/Co-based composite cladding, Micro-hardness, Co-based self-fluxing alloy, Hot Die SKD61 steel

## I. INTRODUCTION

During the last decades, Laser Cladding has known one of the advanced methods to improve the high-quality of surface metal, it is grown with the breakneck speed and particularly advantageous in the repair and restoration the large size parts [1,2]. The melting carbides on based metal techniques are developed strongly to improve and advance wear resistance of the metal. Inside, the Coban self-melting [3] cladding is able to advance oxidation and wear resistance, at the same time, adding more types of carbides such as WC, B<sub>4</sub>C, SiC, Cr<sub>2</sub>C<sub>3</sub>, TiC...with high thermal stability on Co-based to improve mechanical properties at high thermal effectively. Besides, the distribution of TiC particles (%wt) in the clad has a direct effect on cladding hardness in improving work conditions of the hot die's surface. However, the early studies had only researched the microstructure and mechanical properties of cladding.

From those theories, the authors created and studied Co+TiC composite claddings about analyzing, comparing the affection of TiC particles content (%wt) to the hardness of claddings.

## II. EXPERIMENTAL & TESTING METHOD

### A. Experimental method

The experiment is performed on the surface of SKD61 steel, with the main component after Energy Dispersion Analysis (EDS), which is shown in Table 1, with size: 100mm x 30mm x 10mm. Before the experiment, the surface of the samples are cleaned by abrasive paper, washed by alcohol and acetone and then dried by a kiln. Co50 self-melting alloy powder compositions have included: 0.6%C, 3%W, 3.5%Si, 2.25%B, 20%Cr, 5.1% Mo, 5%Fe, 14%Ni and extant of %Co, particles size ~53μm; the fine TiC powder is 99.5% with particles size ~10μm. The composite powders were covered on the SKD61 surface and then used a laser to melt the previous cladding layers. The composite powders are (10%; 20%; 30% wtTiC) + Co50, every thickness is ~1 mm and dried by Kiln in 8 hours before being melted by laser. The clad on SKD61 surface process was performed at Kunming Polytechnic University, on Laser Cladding type GS - TFL 6000 transverse - flowCO<sub>2</sub>, with Laser power is from 3.3 to 3.9 kW, scanning speed is from 350 to 400 mm/min, distance from Laser box to SKD61 surface is 50 mm, Argon flow is 8 L/h, the parameters are shown in Table 2.

TABLE I. CHEMICAL COMPOSITION (% WEIGHT) OF SKD61 STEEL USED IN THE EXPERIMENT.

Element	C	Si	Mn	Cr	Mo	V	Fe
% Weight	0.43	1.17	0.48	4.79	1.38	0.94	extant

TABLE II. EXPERIMENTAL PARAMETERS ARE APPLIED IN LASER CLADDING

Sample	Coating rate (Co: TiC)	Laser power P (kW)	Speed Vs(mm/min)
S1	9 : 1	3.6	500
S2	4 : 1	3.9	400
S3	7 : 3	4.2	450

### B. Mechanical Testing Method

The experiment used the metallographic cutter to minimize the damage by heating; OmniMet electronic scanner and specialized software using halogen light source to observe the microstructure; The Vickers Tukon 1102 micro-hardness tester, with a load of 1,961N (HV0.2). The experiment used the coordinate system as shown in Fig. 1 with the positive direction by the cladding height, and then proceeded to measure the hardness both cladding zone and bonding zone.

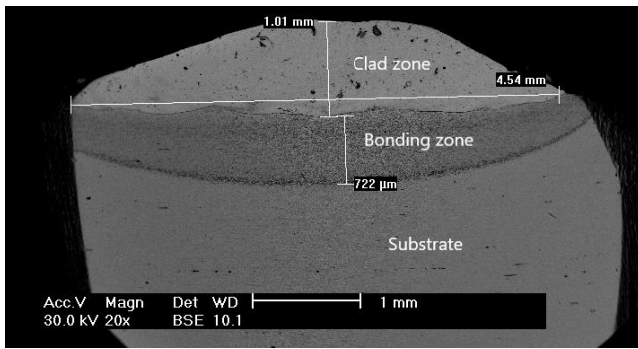


Fig. 1. The experimental coordinate system

### III. RESULTS

#### A. The cladding hardness

The hardness distribution of claddings is shown in Fig. 2. It is easy to see that the hardness is divided into three specific areas, the hardness value of each region is relatively high. From the substrate to the top of clad, the hardness is increasing, which means the wearing resistance of SKD61 steel surface also increases accordingly.

TABLE III. THE RESULTS OF THE MICRO-HARDNESS

Distance from bonding line (μm)	SAMPLE			
	S0	S1	S2	S3
-1400	194.4	235.2	237.5	210.6
-1200	222	241.5	239.5	216.9
-1000	285.4	249.6	244.8	219.9
-800	416.1	483.5	432	459
-600	579.9	683.5	529.5	614.1
-400	567.2	614.6	687	619.4
-200	584.1	566.6	724.5	697.8
0	590.3	552.7	755.3	790.5
200	614.6	618.6	816.3	952.8
400	629.1	568.1	794.6	1038.8
600	635.4	594.6	686.4	935.6

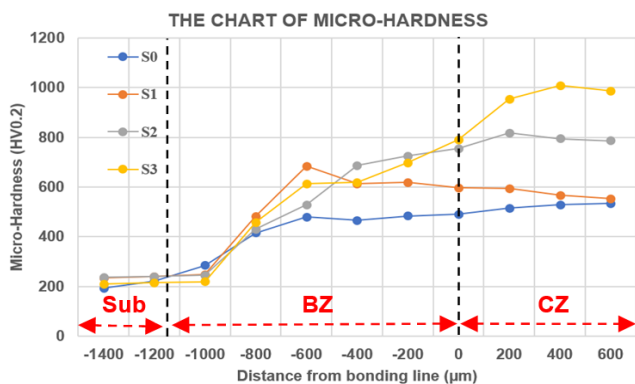


Fig. 2. Variation of microhardness with vertical depth from the top surface of Laser Cladding.

(Sub: Substrate, BZ: Bonding zone, CZ: Clad zone)

From the experimental results, the average hardness of SKD61 heat-treated is 224.8 HV0.2, the average hardness of samples S1, S2 and S3 is 589 HV0.2, 788.2 HV0.2 and 934.4 HV0.2, respectively. All of these claddings have a higher hardness than SKD61 steel, 2.6, 3.5, and 4.2 times, respectively. This result proves that the surface hardness SKD61 steel is significantly improved when applying Laser Cladding method. The bonding zone between the coating and the steel substrate has a lower hardness than the coating, which is due to the presence of Fe from the diffused steel substrate over the coating during the melting process under the effect of heat from a large power laser beam, these Fe atoms have a bad effect of diluting other components, resulting in reduced bond hardness. The heat-affected zone has a higher hardness than the substrate, the main reason is that the alloying elements Co, Mo and other elements, also under the action of the laser beam, make them diffuse to the steel substrate form an enhanced solid solution, another reason is the affection of laser power which causes the temperature of the heat-affected zone to be higher than Ac3 critical point, and then the temperature drops suddenly, which causes the heat-affected zone is quenched (same as the transformation Austenite process when high cooling), causing this zone hardness is quite high [4]. From farther bonding zone, the temperature lower, the difference between the non-heated and heated smaller, eventually the heated areas are overrated to the organization of the steel base, so the hardness also decreases [5]. The S1, S2 and S3 cladding have a gradual increase in hardness. As the TiC content in the composite cladding component increases, the hardness also increases with the corresponding, the main causes are the newly formed phases with high hardness and melting point, and the individual atoms of Ti and C also react to recombinate forming the new TiC is more complete than before. In Addition, the cladding also contains the non-molten TiC and a partial melting TiC that retains the main characteristics of the carbides, which are small in size and evenly dispersed so that the hardness of these claddings increases.

#### B. Modelling the distribution of TiC reinforced Co-based

From the results in Table 4, the experiment uses data in "Objects' Area" cell in the result column to create the graphs showing the distribution (%) TiC particles along with the height of sample S1, S2, S3 clad.

The regression model was found by two methods, Microsoft Excel software (The Linear line on Fig. 3,4,5) and Parabolic interpolation method [6]. With the significance level selected  $p = 5\%$  the empirical data is in the allowable error of the expression; therefore, the linear equation showing the regression model of TiC particle distribution along the height of the newly discovered coatings is suitable and shown in Fig. 3, 4, 5.

From there describe the expressions as follows:

$$S1: y_1 = -7.3494x + 23.8438$$

$$S2: y_2 = 9.7389x + 44.6927$$

$$S3: y_3 = 13.6597x + 35.511$$

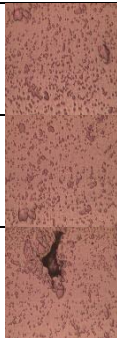
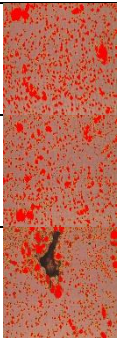
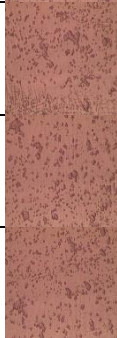
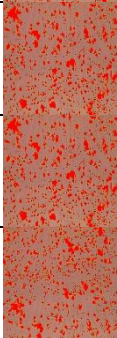
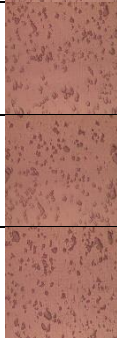
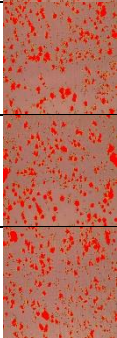
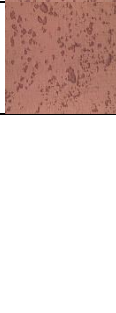

The expression of S1 pattern tends to go down this means that the further the TiC particle distribution ratio decreases, the further away from the bonding surface. With the solution mentioned in the study is to improve the hot stamping surface. Still, in the S1 sample, the closer the TiC surface is to the cladding surface, the more the TiC particle distribution ratio decreases significantly, so S1 will not improve much the



surface quality of hot die. The reason is the TiC particles content in the small coating component (10% weight) so under the effect of a laser beam, a part of TiC is completely melted and resolved into [Ti] atoms and [C], these new [Ti] atoms react with the element Co available in the composition of Co50, forming the new TiCo3 phase [7].

The expression of the sample group S2 and S3, the graph is sloping up, which means that the farther the surface between the steel substrate and the coating surface, the higher the TiC particle distribution ratio will be, contributing to a significant improvement in the amount of hot stamping surface that actual demand is posing. Since the TiC particle content in the cladding composition is significant (20% weight in sample S2 and 30% in the S3 coating), when there is a laser heat source, the dilution forms and with the mechanism, the quick cooling from the outside to the perpendicular direction of the heat source will cause the TiC particles above the cladding surface to remain in place, while the TiC particles in the middle of the cladding will have a slower cooling speed, so tendency to diffuse down the bonding line.

TABLE IV. THE RESULTS OF THE DISTRIBUTION TiC PARTILES ALONG HEIGHT OF SAMPLE S1 CLAD

Distance from bonding line (mm)	Etched	Analysed	Result
0.125			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 1.827</div> <div>Range: 1.827</div> <div>Sum: 24.439</div> <div>Mean: 0.017</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.000</div> <div>Total Area (mm²): 1338.834</div> <div>Cladding Area (%): 24.439</div> <div>Cladding: 839</div>
0.25			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 1.097</div> <div>Range: 1.097</div> <div>Sum: 27.175</div> <div>Mean: 0.027</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.011</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 27.175</div> <div>Cladding: 753</div>
0.375			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.947</div> <div>Range: 0.947</div> <div>Sum: 26.849</div> <div>Mean: 0.023</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.016</div> <div>Total Area (mm²): 1258.834</div> <div>Cladding Area (%): 26.849</div> <div>Cladding: 837</div>
0.5			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.982</div> <div>Range: 0.982</div> <div>Sum: 18.988</div> <div>Mean: 0.027</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.016</div> <div>Total Area (mm²): 1275.881</div> <div>Cladding Area (%): 18.988</div> <div>Cladding: 655</div>
0.625			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.982</div> <div>Range: 0.982</div> <div>Sum: 18.988</div> <div>Mean: 0.027</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.016</div> <div>Total Area (mm²): 1275.881</div> <div>Cladding Area (%): 18.988</div> <div>Cladding: 655</div>
0.75			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 2.804</div> <div>Range: 2.804</div> <div>Sum: 17.880</div> <div>Mean: 0.026</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.016</div> <div>Total Area (mm²): 1288.834</div> <div>Cladding Area (%): 17.880</div> <div>Cladding: 583</div>
0.875			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.196</div> <div>Range: 0.196</div> <div>Sum: 36.475</div> <div>Mean: 0.004</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.001</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 36.475</div> <div>Cladding: 683</div>
1			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.557</div> <div>Range: 0.557</div> <div>Sum: 34.621</div> <div>Mean: 0.004</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.001</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 34.621</div> <div>Cladding: 783</div>
1.125			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.577</div> <div>Range: 0.577</div> <div>Sum: 31.855</div> <div>Mean: 0.003</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.001</div> <div>Total Area (mm²): 1225.834</div> <div>Cladding Area (%): 31.855</div> <div>Cladding: 625</div>
1.25			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.589</div> <div>Range: 0.589</div> <div>Sum: 31.175</div> <div>Mean: 0.003</div> <div>Variance: 0.000</div> <div>Std. Dev: 0.001</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 31.175</div> <div>Cladding: 585</div>

Compared with the sample graph S3, the sample graph S2 has a smaller slope, and there is no big difference in the TiC particles distribution ratio along with the cladding height. In the sample S3, there appeared cracks in the bonding layer, due to the relatively high TiC grain content along with the selection of unreasonable technological parameters, so the unformed cladding was strongly bonded the steel surface substrate. From the observations and observations above, we can see the sample graph S2 with the most suitable regression model in all samples selected for the experiment.

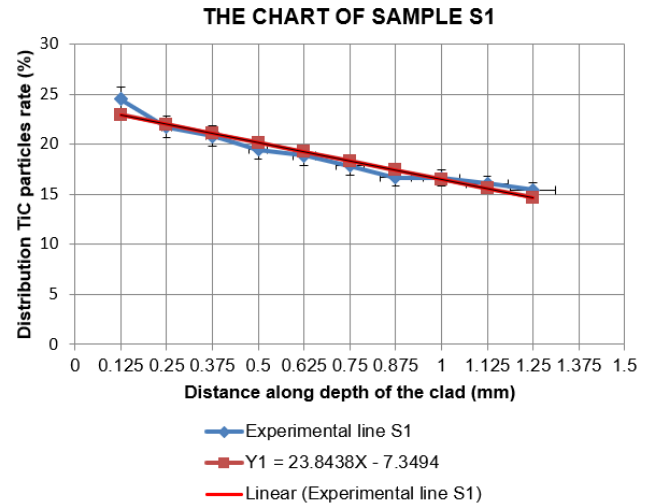
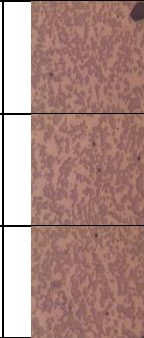
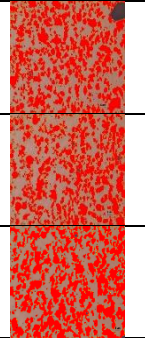

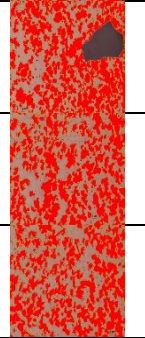
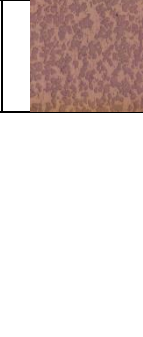
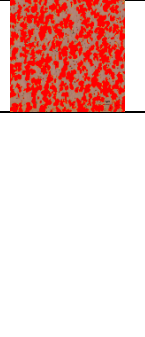




Fig. 3. Modelling the distribution of TiC in S1

TABLE V. THE RESULTS OF THE DISTRIBUTION TiC PARTILES ALONG HEIGHT OF SAMPLE S2 CLAD

Distance from bonding line (mm)	Etched	Analysed	Result
0.125			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 3.391</div> <div>Range: 3.391</div> <div>Sum: 42.678</div> <div>Mean: 0.35</div> <div>Variance: 0.188</div> <div>Std. Dev: 0.433</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 42.678</div> <div>Cladding: 722</div>
0.25			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 1.923</div> <div>Range: 1.923</div> <div>Sum: 41.174</div> <div>Mean: 0.168</div> <div>Variance: 0.118</div> <div>Std. Dev: 0.338</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 41.174</div> <div>Cladding: 704</div>
0.375			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 0.524</div> <div>Range: 0.524</div> <div>Sum: 50.527</div> <div>Mean: 0.100</div> <div>Variance: 0.073</div> <div>Std. Dev: 0.270</div> <div>Total Area (mm²): 1238.834</div> <div>Cladding Area (%): 50.527</div> <div>Cladding: 620</div>
0.5			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 23.242</div> <div>Range: 23.242</div> <div>Sum: 22.482</div> <div>Mean: 0.367</div> <div>Variance: 7.888</div> <div>Std. Dev: 2.808</div> <div>Total Area (mm²): 1238.834</div> <div>Cladding Area (%): 22.482</div> <div>Cladding: 504</div>
0.625			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 6.751</div> <div>Range: 6.751</div> <div>Sum: 49.827</div> <div>Mean: 0.121</div> <div>Variance: 0.413</div> <div>Std. Dev: 0.643</div> <div>Total Area (mm²): 1238.834</div> <div>Cladding Area (%): 49.827</div> <div>Cladding: 774</div>
0.75			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 11.898</div> <div>Range: 11.898</div> <div>Sum: 48.47</div> <div>Mean: 0.120</div> <div>Variance: 0.194</div> <div>Std. Dev: 0.441</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 48.47</div> <div>Cladding: 722</div>
0.875			<div>Statistics</div> <div>% Area</div> <div>Min: 0</div> <div>Max: 8.686</div> <div>Range: 8.686</div> <div>Sum: 34.251</div> <div>Mean: 0.168</div> <div>Variance: 0.819</div> <div>Std. Dev: 0.905</div> <div>Total Area (mm²): 1705.881</div> <div>Cladding Area (%): 34.251</div> <div>Cladding: 297</div>

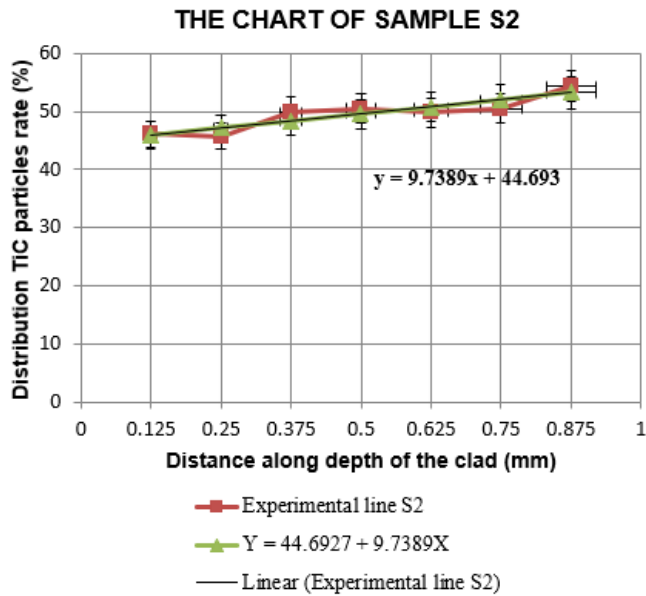


Fig. 4. Modelling the distribution of TiC in S2

TABLE VI. THE RESULTS OF THE DISTRIBUTION TiC PARTILES ALONG HEIGHT OF SAMPLE S3 CLAD

Distance from bonding line (mm)	Etched	Analysed	Result
0.125			Statistics % Area Min: 0 Max: 1.41 Range: 1.41 Sum: 58.249 Mean: 0.006 Variance: 0.002 Std.Dev: 0.043 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 38.234 Precision: 0.76
0.25			Statistics % Area Min: 0 Max: 2.81 Range: 2.81 Sum: 30.123 Mean: 0.003 Variance: 0.002 Std.Dev: 0.043 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 37.725 Precision: 0.82
0.375			Statistics % Area Min: 0 Max: 2.603 Range: 2.603 Sum: 41.112 Mean: 0.008 Variance: 0.002 Std.Dev: 0.043 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 43.432 Precision: 0.88
0.5			Statistics % Area Min: 0 Max: 2.374 Range: 2.374 Sum: 42.112 Mean: 0.006 Variance: 0.003 Std.Dev: 0.039 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 41.112 Precision: 0.74
0.625			Statistics % Area Min: 0 Max: 2.527 Range: 2.527 Sum: 42.247 Mean: 0.006 Variance: 0.003 Std.Dev: 0.039 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 43.217 Precision: 0.74
0.75			Statistics % Area Min: 0 Max: 3.149 Range: 3.149 Sum: 40.271 Mean: 0.003 Variance: 0.003 Std.Dev: 0.039 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 41.112 Precision: 0.83
0.875			Statistics % Area Min: 0 Max: 4.372 Range: 4.372 Sum: 47.242 Mean: 0.008 Variance: 0.003 Std.Dev: 0.039 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 47.242 Precision: 0.88
1			Statistics % Area Min: 0 Max: 3.149 Range: 3.149 Sum: 40.271 Mean: 0.003 Variance: 0.003 Std.Dev: 0.039 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 41.112 Precision: 0.83
1.125			Statistics % Area Min: 0 Max: 2.589 Range: 2.589 Sum: 49.837 Mean: 0.008 Variance: 0.003 Std.Dev: 0.039 Total Area (mm <sup>2</sup> ): 12505.894 Measured Area (%): 49.837 Precision: 0.88

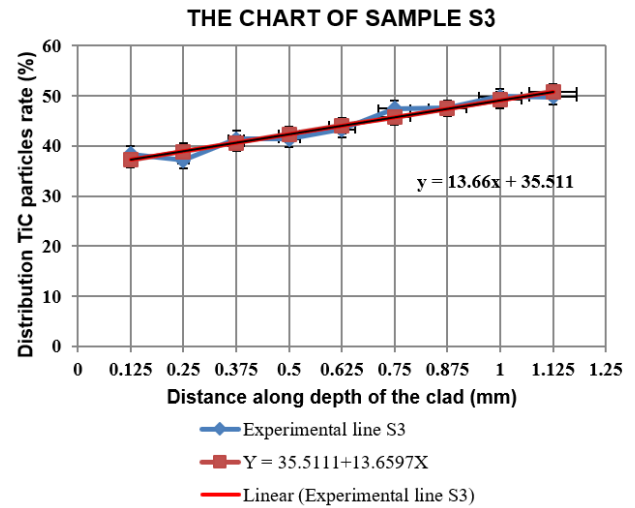


Fig. 5. Modelling the distribution of TiC in S3

#### IV. CONCLUSION

The cladding hardness is significantly improved, the average hardness of the cladding Co+10%Ti, Co+20%TiC and Co+ 0% TiC are 589 HV0.2, 788.2 HV0.2 and 934.4 HV0.2. All of these claddings have higher hardness than SKD61 steel respectively 2.6, 3.5 and 4.2 times, thereby showing that the cladding formed by Laser Cladding method is completely possible to improve the surface quality of hot stamping mold.

The modelling of TiC particle distribution for each sample with concentrations of 10%, 20% and 30% TiC are:  $y_1 = 23.8438 - 7.3494x$ ,  $y_2 = 9.7389x + 44.6927$ ,  $y_3 = 13.6597x + 35.511$ , completely consistent with the results of hardness measurement of claddings. Thereby, if building regression models of particle distribution carbides can accurately predict coating hardness, creating a theoretical basis to apply to other types of reinforced carbides others.

#### REFERENCE

- [1] Hardro J P, "Development of materials for the rapid manufacture of die cast tooling," United States - Newport: Graduate University of Rhode Island, 2001.
- [2] Chen J Y, Conlon K, Xue L, et al., "Experimental study of residual stresses in laser clad AISI P20 tool steel on pre-hardened wrought P20 substrate, Materials Science and Engineering," 2010, Vol. 527, No. 27-28, pp. 7265-7273.
- [3] Zhao Y M, Wang J L, Mou J W, "Microstructures and properties of Co-based alloy claddings prepared on surface of H13 steel," China Welding, 2010, Vol. 19, No. 3, pp. 41-44.
- [4] Qian X Y, Tong H Q, Zhang D L, et al, "Microstructure and performance of laser-cladding Co-based alloy cladding on the surface of H13 mold steel," Metallurgical Collections, 2011, Vol. 5, pp. 1-3. (in China).
- [5] Si S H, He Y ZH, Yuan X M, et al, "Microstructure and wear-resistance of laser clad Co-based alloy claddings with B4C3 and SiC3," The Chinese Journal of Nonferrous Metals, 2003, Vol. 13, No. 2, pp. 454-459. (in China).
- [6] Phung Ran, "Experimental Planning, Ho Chi Minh City University of Technology and Education," 2015.
- [7] Pham Thi Hong Nga, Tran The San, Jiang Ye-hua, "Microstructure and Mechanical property of TiC/Co composite cladding on the AISI H13 surface by Laser Cladding methods," Viet Nam Mechanical Journal, 2013, Vol. 9, pp. 58-63.

# Real-time Measurement and Prediction of Ball Trajectory for Ping-pong Robot

Vo Duy Cong

Industrial Maintenance Training Center  
Ho Chi Minh city University of  
Technology, VNU-HCM  
Ho Chi Minh city, Viet Nam  
congvd@hcmut.edu.vn

Le Duc Hanh

Faculty of Mechanical Engineering  
Ho Chi Minh city University of  
Technology, VNU-HCM  
Ho Chi Minh city, Viet Nam  
ldhanh@hcmut.edu.vn

Le Hoai Phuong

Industrial Maintenance Training Center  
Ho Chi Minh city University of  
Technology, VNU-HCM  
Ho Chi Minh city, Viet Nam  
lhphuong@hcmut.edu.vn

**Abstract**— This paper develops a vision system to predict ball trajectory for the ping-pong robot. The stereo vision system with two cameras is used to measure ball center. A multi threshold segmentation algorithm is applied to precisely locate the center of ball in image and the triangulation algorithm is applied to compute the 3D position in the world coordinates. Six consecutive coordinates are collected and the least squares method is used to determine the initial states of the ball. From these initial states, the following flight trajectory of the ball is predicted using two physical models: aerodynamics model and rebound model. Then the states of the ball at striking point are calculated from the predicted trajectory. Experimental results show that the developed system can achieve a good predicting precision in real-time.

**Keywords**—Ping pong robot, stereo vision, visual measurement, trajectory prediction, triangulation algorithm.

## I. INTRODUCTION

The table tennis robot has become a hot topic and has been attracting many researchers because of the difficulty and challenges of vision, real-time and intelligent control. The first table tennis robot was developed by Andersson at AT&T Bell Laboratories in 1988 [1]. After that, many research groups have focused on developing robot systems both in mechanical structure, vision system and control [2]. The vision system is a crucial component, the accuracy and latency of image processing due to the low sampling rate of vision system directly affect the control performance of the robot system.

The vision system, like the human eye, is used to determine the 3D position and velocity of the ball. From that, the trajectory of the ball and robot can be predicted. Therefore, the accuracy of the vision system is very important, it decides to the accuracy of the prediction and control performance. The vision system consists of cameras and hardware used for image processing (PC or special processor). Table I lists some vision systems developed since 1988.

TABLE I. DEVELOPMENT OF THE VISION SYSTEM FOR TABLE TENNIS ROBOTS

Year	Author	Number of cameras, speed, processor
1988	Anderson [1]	4 camera, 60Hz, MC68020 processor
1990	Fassler [5]	2 camera, 50Hz, MC68000 processor
2003	L. Acosta [6]	1 camera, 40Hz, PC
2005	Miyazaki [3]	2 camera, 60Hz, Quick MAG
2005	K.P. Modi [7]	1 USB camera, 15Hz, work station

Year	Author	Number of cameras, speed, processor
2005	Y. Zhang	1 camera, 60-89Hz, PC
2006	Y. Zhang	2 cameras, 60-89Hz, PC
2007	Quanta-View Inc	2 cameras 60Hz, Intel Xeon processors
2010	Z. Zhang [8]	2 cameras (DSP, FPGA), 250Hz, PC
2012	Li, Hailing [9]	2 cameras 150Hz, 2 PC

The number of cameras can be one, two or more. Using more cameras can improve the accuracy and can help robot operate in many different environments. However, the system becomes more complicated and the processor must have high performance to simultaneously process multiple frames and synchronization. In addition, the camera calibration is also more complex. Using one camera, only a single frame is needed to be processed, so the real-time performance was improved. Acosta et al. [6] used a monocular vision system, which compute the 3-D position of the ball according to the image coordinates of the ball and its shadow on the table and so it required a stable light-controlled environment.

Nowadays, the cameras have very fast shutter speed. However, the ineffectiveness of image processing algorithms for ball recognition and tracking will lead to reduced performance of vision system. Sabzevari et al. [10] and Lampert et al. [11] both developed vision system which only use color features to detect the ball. Only using color features can lead to wrong detection, because other objects are easily to be recognized as the orange ball if they are the same color as the ball. In [8], the flying ball can be segmented by using adjacent frame difference. This segmentation method might often obtain incomplete ball contours and cannot detect the table tennis ball when the ball flies slowly.

To return the ball at a certain area on the region of opponent, the robot must control the paddle to hit the ball with a required velocity. So, the robot must have a motion planning to achieve the desired velocity at the striking point. The position and velocity of paddle at striking point are determined from the position and velocity of the ball. For this reason, the control system needs to predict trajectory of the ball and the hitting position will be estimated from this prediction trajectory. Zhang et al [8] used two physical models to construct the iteration for prediction.

In this paper, a vision system is developed to measure and predict trajectory of the ball for the ping-pong robot. A binocular vision system captures and transmits the images to a PC. On the PC, a multi threshold segmentation algorithm is



applied to find the center of the ball in two images. The 3D position is computed using the triangulation algorithm that is widely used in stereo vision systems. The aerodynamics model and rebound models of the ball are established to estimate and predict its trajectory.

The remainder of this paper is organized as follows: Section II describes the algorithms used in the vision system. Section III discusses how to predict the trajectory using two physical models. Experiment results are shown in section IV.

## II. VISON SYSTEM

The goal of the vision system is to determine position of the ball for predicting the trajectory. Therefore, the accuracy and high processing speed of the vision system are required to achieve the high-performance trajectory prediction. In this section, a stereo vision system with a set of effective ball recognition and tracking algorithms is proposed to measure the ball center.

### A. Determine the position of the ball in the image

To ensure the accuracy and processing speed, it is necessary to develop an optimal image processing algorithm. This section uses multi threshold segmentation algorithm with the cooperation of HSV threshold, Subtract Background method and filters based on the size, shape or area of the ball to remove objects that have the same color as the ball.

1) *Subtract Background*: Use a known background image without the ball. The difference in gray level of the current image and the background image will extract the ball from background, i.e.,

$$I_{diff}(u, v) = I_c(u, v) - I_b(u, v) \quad (1)$$

where  $I_c(x, y)$  is the gray value of pixel  $(u, v)$  in the current image,  $I_b(u, v)$  is the gray value in base image and  $I_{diff}(u, v)$  is the gray value of pixel  $(u, v)$  in the image formed by the frame difference. After the frame difference threshold, the image is binarized:

$$Diff(u, v) = \begin{cases} 1, & \text{if } |I_{diff}(u, v)| > Gray_{Thresh} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

2) *HSV Threshold*: Because the light is not stable, only using frame difference threshold, the detection sometimes will be faulty. Therefore it is necessary to combine with the color threshold. After convert BGR image to HSV image, the image is thresholded using six range thresholds  $Hue_{min}$ ,  $Hue_{max}$ ,  $Sat_{min}$ ,  $Sat_{max}$ ,  $Val_{min}$  and  $Val_{max}$  according to the H, S, V features of uniform orange table tennis ball. These values will be determined based on experiments.

Apply the Subtract Background and HSV Threshold methods, we obtain two binary images. The final image is the sum of these two images.

3) *Threshold using shape features*: Image after applying HSV threshold and Subtrack Background will contain objects with the same color as the ball or noise due to lighting conditions as shown in Fig 1.

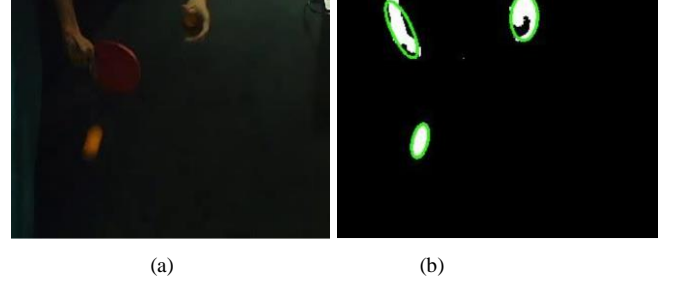


Fig. 1. Process image. (a) image capture from camera, (b) Threshold image using Subtract Background and HSV threshold.



Fig. 2. Contour image after applying the shape thresholds

To eliminate these noise objects, perform a contour finding of the objects, then filter the noise based on radius and area of contour:

$$R_{max} \geq r_c \geq R_{min} \quad (3)$$

$$Area_{max} \geq a_c \geq Area_{min} \quad (4)$$

where  $r_c$  and  $a_c$  are the radius and area of ellipse interpolated from object contour.  $R_{max}$ ,  $R_{min}$ ,  $Area_{max}$ ,  $Area_{min}$  are the shape feature thresholds.

Using additional shape feature threshold has significantly improved the robustness of the ball detection algorithm. The result is shown in Fig 2.

### B. Determine 3D position

Fig 3 shows the camera system used in the robot system, in which two cameras are placed perpendicular to each other. A coordinate system  $\Sigma_B$  located in the corner of the table is used as the reference of the system, the ball coordinates will be determined in this coordinate system. The camera calibration will determine the relationship between the two camera coordinates  $\Sigma_{C1}$  and  $\Sigma_{C2}$  vs  $\Sigma_B$ .

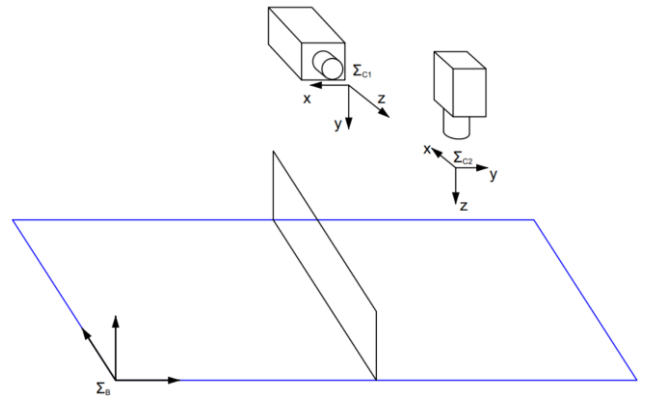


Fig. 3. Camera system



Using the algorithm described above, two the centers of the ball in the two images are determined. The 3D coordinate of the ball is computed based on the idea of a triangulation method that is widely used in stereo vision systems. However, instead of reconstructing the entire 3D space, we only need to determine the coordinate of the ball center, so the process is simpler and ensures the real-time.

Denote the 3D coordinate of the ball is  ${}^B P$  in  $\Sigma_B$  and  ${}^{C_n} P$  in  $\Sigma_{C_n}$  ( $n = 1, 2$ ). We have:

$${}^{C_n} P = {}^{C_n} P_B + {}^{C_n} R_B {}^B P \quad (5)$$

where  ${}^B P = (X_B, Y_B, Z_B)^T$ ,  ${}^{C_n} P = (X_{C_n}, Y_{C_n}, Z_{C_n})^T$  and  ${}^{C_n} P_B \in R^3$  is the coordinate of the origin of  $\Sigma_B$  in  $\Sigma_{C_n}$ ,  ${}^{C_n} R_B \in R^{3 \times 3}$  is the rotation matrix from  $\Sigma_B$  to  $\Sigma_{C_n}$ :

$${}^{C_n} P_B = \begin{bmatrix} X_{BC_n} \\ Y_{BC_n} \\ Z_{BC_n} \end{bmatrix} \quad {}^{C_n} R_B = \begin{bmatrix} r_{11}^n & r_{12}^n & r_{13}^n \\ r_{21}^n & r_{22}^n & r_{23}^n \\ r_{31}^n & r_{32}^n & r_{33}^n \end{bmatrix}, n = 1, 2 \quad (6)$$

From (5), (6) and the formula of the pinhole camera model, perform several mathematical transformations, we derive the matrix equation:

$$A \cdot {}^B P = B \quad (7)$$

with:

$$A = \begin{bmatrix} \frac{u_1}{f_x} r_{31}^1 - r_{11}^1 & \frac{u_1}{f_x} r_{32}^1 - r_{12}^1 & \frac{u_1}{f_x} r_{33}^1 - r_{13}^1 \\ \frac{v_1}{f_y} r_{31}^1 - r_{21}^1 & \frac{v_1}{f_y} r_{32}^1 - r_{22}^1 & \frac{v_1}{f_y} r_{33}^1 - r_{23}^1 \\ \frac{u_2}{f_x} r_{31}^2 - r_{11}^2 & \frac{u_2}{f_x} r_{32}^2 - r_{12}^2 & \frac{u_2}{f_x} r_{33}^2 - r_{13}^2 \\ \frac{v_2}{f_y} r_{31}^2 - r_{21}^2 & \frac{v_2}{f_y} r_{32}^2 - r_{22}^2 & \frac{v_2}{f_y} r_{33}^2 - r_{23}^2 \end{bmatrix}$$

$$B = \begin{bmatrix} X_{BC_1} - \frac{u_1}{f_x} Z_{BC_1} \\ Y_{BC_1} - \frac{v_1}{f_y} Z_{BC_1} \\ X_{BC_2} - \frac{u_2}{f_x} Z_{BC_2} \\ Y_{BC_2} - \frac{v_2}{f_y} Z_{BC_2} \end{bmatrix}$$

The solution  ${}^B P$  of equation (7) can be derived using Least Square Error (LSE) method:

$${}^B P = A^T B \quad (8)$$

where  $A^T = (A^T A)^{-1} A^T$  is the pseudo-inverse matrix of matrix  $A$ .

### III. BALL TRAJECTORY PREDICTION

Predicting the trajectory of the ball is the first important task of control system to determine the hitting point and velocity of ball at the hitting position. These parameters are necessary to determine the required position, velocity and rotation of racket fixed on the robot arm and then the trajectory of racket can be planned to return the ball. Fig 4 depicts the trajectory prediction process since opponent hits the ball until the robot returns the ball.

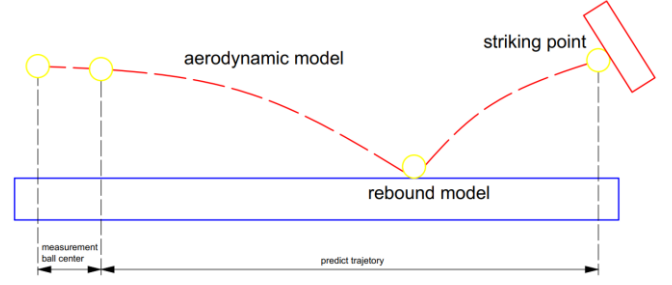


Fig. 4. Ball trajectory prediction.

It can be seen in Fig 4, to predict ball trajectory, we need to use two models: aerodynamics model and table rebound model.

#### A. Aerodynamics model

When the ball moves through the air, there are four forces on the ball, include: the Magnus Force  $F_m$ , gravity  $F_g$ , air resistance  $F_d$  and buoyancy force  $F_b$  [12]:

$$F_d = -\frac{1}{8} \rho_a \pi D^2 C_D \|v\| v \quad (9)$$

$$F_m = \frac{1}{8} C_m \rho_a \pi D^3 \omega \times v \quad (10)$$

$$F_g = mg \quad (11)$$

$$F_b = m_b g \quad (12)$$

where  $\rho_a$  is air density,  $D$  is the diameter of the ball,  $C_D$  is the drag coefficient and  $v$  is the velocity of the ball.  $\omega$  is the angular velocity,  $m$  is the mass of the ball,  $m_b$  is the mass of the air with the same volume as the ball,  $g$  is the gravity accelerator and  $C_m$  is the Magnus coefficient. According to the standard rules, the mass of the ball  $m = 0.00275 \text{ kg}$ , ball diameter  $D = 0.04 \text{ m}$ , gravity accelerator  $g = 9.81 \text{ m/s}^2$ ,  $\rho_a = 1.29 \text{ kg/m}^3$ .

Because  $m_b \ll m$ ,  $F_b$  will be ignored and not considered in the model. If the ball rotates with large angular velocity, the Magnus force can reach the same level as the air resistance and cannot neglect. However, it is difficult to measure the angular velocity of the ball for the vision system, so in this paper assumes the ball fly without rotation. Apply Newton's second law  $\sum F = m\dot{v}$ , derive the differential equation:

$$\dot{v} = -g - k_d \|v\| v \quad (13)$$

where  $k_d = C_D \rho_a \pi D^2 / 8m$ .

We know that velocity is the derivative of position:

$$\dot{v} = \dot{p} \quad (14)$$

Combine (13) and (14), the motion of the ball can be represented by state vector equation:

$$\begin{bmatrix} \dot{p}_x \\ \dot{p}_y \\ \dot{p}_z \\ \dot{v}_x \\ \dot{v}_y \\ \dot{v}_z \end{bmatrix} = \begin{bmatrix} v_x \\ v_y \\ v_z \\ -k_d \|v\| v_x \\ -k_d \|v\| v_y \\ -k_d \|v\| v_z - g \end{bmatrix} \quad (15)$$

### B. Rebound model

After the ball bounces off the table, the velocity of the ball will be changed in both direction and magnitude. Denote  $(v_b, \omega_b)$  and  $(v'_b, \omega'_b)$  are the velocity and the angular velocity of the ball before and after bouncing. In fact, the velocity after bouncing is changed due to friction and the ball being spin. Define the velocity of the contact point as follows:

$$v_{bT} = [v_{bTx} \quad v_{bTy} \quad 0]^T + \omega_b \vec{r} = \begin{bmatrix} v_{bx} - \omega_{by}r \\ v_{by} + \omega_{bx}r \\ 0 \end{bmatrix} \quad (16)$$

Using rebound model in [13], the velocity of the ball after rebound is determined by equations:

$$\begin{cases} v'_b = V_{vv}v_b + V_{v\omega}\omega_b \\ \omega'_b = V_{\omega v}v_b + V_{\omega\omega}\omega_b \end{cases} \quad (17)$$

There are two cases for the coefficient matrices:

*Case 1:* the velocity of the contact point is not zero:

$$V_{vv} = \begin{bmatrix} 1-\delta & 0 & 0 \\ 0 & 1-\delta & 0 \\ 0 & 0 & -e_t \end{bmatrix}, \quad V_{v\omega} = \begin{bmatrix} 0 & \delta r & 0 \\ -\delta r & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$V_{\omega v} = \begin{bmatrix} 0 & -\frac{3\delta}{2r} & 0 \\ \frac{3\delta}{2r} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad V_{\omega\omega} = \begin{bmatrix} 1-\frac{3\delta}{2} & 0 & 0 \\ 0 & 1-\frac{3\delta}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where  $\delta$  is equal to  $\mu(1+e_t)|v_{bz}|/\|v_{bT}\|$ . This situation occurs when the condition (18) is satisfied:

$$\frac{\mu(1+e_t)|v_{bz}|}{\|v_{bT}\|} \leq 0.4 \quad (18)$$

*Case 2:* If (18) is not satisfied, the velocity of the lowest point is zero and:

$$V_{vv} = \begin{bmatrix} 0.6 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & -e_t \end{bmatrix}, \quad V_{v\omega} = \begin{bmatrix} 0 & 0.4r & 0 \\ -0.4r & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$V_{\omega v} = \begin{bmatrix} 0 & -\frac{0.6}{r} & 0 \\ \frac{0.4}{r} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad V_{\omega\omega} = \begin{bmatrix} 0.4 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

### C. Trajectory prediction

In order to predict the trajectory, the coefficients and initial parameters need to be determined. A series of discrete 3-D positions with timestamp on the trajectory are obtained from stereo vision system. These data contain many errors caused by image processing or camera calibration and so on. Thus, the second order polynomials are used to fit vision data to reduce the noise:

$$\begin{cases} x = a_1t^2 + b_1t + c_1 \\ y = a_2t^2 + b_2t + c_2 \\ z = a_3t^2 + b_3t + c_3 \end{cases} \quad (19)$$

where  $a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3$  and  $c_3$  are the coefficients of the polynomials for the X, Y and Z coordinates.

These coefficients are easily identified via Least Square Method (LSM). Then, the initial velocity of the ball can be obtained by deriving the equation (19):

$$\begin{cases} v_x = 2a_1t + b_1 \\ v_y = 2a_2t + b_2 \\ v_z = 2a_3t + b_3 \end{cases} \quad (20)$$

Once the coefficients and initial parameters  $[x_0, y_0, z_0, v_{x0}, v_{y0}, v_{z0}]$  are derived, the continued trajectory can be interpolated by expressed equation (15) with iteration:

$$\begin{bmatrix} x_i \\ y_i \\ z_i \\ v_{xi} \\ v_{yi} \\ v_{zi} \end{bmatrix} = \begin{bmatrix} x_{i-1} \\ y_{i-1} \\ z_{i-1} \\ v_{xi-1} \\ v_{yi-1} \\ v_{zi-1} \end{bmatrix} + \begin{bmatrix} v_x \\ v_y \\ v_z \\ -k_d\|v\|v_x \\ -k_d\|v\|v_y \\ -k_d\|v\|v_z - g \end{bmatrix} T \quad (21)$$

where  $i = 1, 2, 3, \dots, T$  is the time interval for one iteration.

When  $z_i$  is equal to the height of table, the ball touches the table at the landing point. The velocity at the landing point obtained from the iteration with (21) serves as the input velocity to compute the velocity of the ball after rebound using rebound model expressed in equation (17). The trajectory after rebound is further predicted using (21) with the initial states are the landing point and the velocity of the ball after rebound. The position of the racket to return the ball is determined when either of the following conditions occurs:

$$\begin{cases} x_i \geq x_{max} \\ v_z \geq 0 \end{cases} \quad (22)$$

When (22) occurs, the position and velocity of the ball will be used to determine the position and velocity of the racket at the striking point.

## IV. EXPERIMENTS

### A. Experiment setup

The vision system uses two cameras run at framerate 50Hz to capture images (648x480 pixels) simultaneously. A chess board is employed to calibrate parameters of two camera. The calibration uses the MATLAB calibration toolbox. The parameters of two cameras after calibration are given in table 2.

TABLE II. CAMERA PARAMETERS.

Parameter	Camera 1			Camera 2		
Intrinsic matrix	311.71	0	163.96	311.22	0	145.78
	0	312.32	108.37	0	318.93	123.9
	0	0	1	0	1	1
Translational vector	[-593.82, -611.08, 1315.72]			[869.84, 361.79, 1402.80]		
Rotational matrix	-0.0232	0.9992	0.0332	-0.9842	-0.1767	-0.0093
	0.997	-0.0228	-0.0112	0.0054	0.0224	-0.9997
	-0.0104	0.0335	-0.994	0.1769	-0.984	-0.0211

### B. Vision system evaluation

Experiments were conducted to evaluate the robustness of image processing algorithms, speed and accuracy of the vision system. The thresholds in the multi threshold segmentation algorithm will be adjusted depending on the environment. In our experiment, the values for  $Hue_{min}$ ,  $Hue_{max}$ ,  $Sat_{min}$ ,  $Sat_{max}$ ,  $Val_{min}$  and  $Val_{max}$  are 0, 40, 120, 255, 86 and 255 respectively. System takes 3-5ms to capture image from the cameras and about 7-10ms to process image and compute 3D coordinate.

The results of using image processing algorithm to detect the ball are shown in Fig 1 and Fig 2 in which the noise object or the object with the same color as the ball was removed. Table 3 shows some 3D coordinate measurement results using the stereo vision system. It can be seen that the computed coordinates are different for each measurement due to the error of image processing. However, this error is only about 2cm which is acceptable.

### C. Trajectory prediction evaluations

1) *Determine the coefficients* : In ADM model,  $k_d$  is determined by experiment. Collect more than forty data, use the LSM method, the  $k_d$  coefficient is estimated about 0.1196.

From equation (17), ignore the angular velocity, the velocity of the ball before and after rebound will relate according to simple formula:

$$v'_b = V_{vv} v_b \quad (23)$$

where

$$V_{vv} = \begin{bmatrix} k_t & 0 & 0 \\ 0 & k_t & 0 \\ 0 & 0 & -e_t \end{bmatrix} \quad (24)$$

Hence, the coefficients  $e_t$  and  $k_t$  need to be determined. The velocities of the ball before and after rebound were computed from the position of the ball according to approximate formula:

$$v_i = \frac{x_i - x_{i-1}}{\Delta t} \quad (25)$$

where the interval  $\Delta t = 20ms$ . Twenty values of velocity are computed and approximate to the first order equation  $v_{out} = av_{in} + b$ . The slope of line is the coefficient that needs to be determined. The estimated coefficients are  $e_t = 0.77$  and  $k_t = 0.634$  and shown in Fig 5 and Fig 6.

TABLE III. VISION MEASUREMENT AND ERRORS.

Index	Actual (x, y, z) mm	Measured (x, y, z) mm	Error (x, y, z) mm
1	[560, 420, 20]	[564.43, 410.56, 16.64]	[4.43, 9.44, 3.36]
2	[560, 420, 20]	[576.12, 428.43, 14.10]	[16.12, 8.43, 5.9]
3	[560, 420, 20]	[578.34, 439.71, 6.05]	[18.34, 19.71, 13.95]
4	[560, 420, 20]	[557.15, 408.65, 23.68]	[2.85, 11.35, 17.15]
5	[324, 335, 20]	[330.73, 346.12, 4.17]	[6.73, 11.12, 13.27]
6	[324, 335, 20]	[328.19, 316.15, 39.45]	[4.19, 18.85, 15.81]
7	[324, 335, 20]	[316.84, 337.46, 39.86]	[7.16, 2.46, 19.86]
8	[324, 335, 20]	[317.66, 324.27, 28.93]	[6.34, 10.73, 8.93]
9	[68, 127, 110]	[74.45, 128.96, 106.18]	[6.45, 1.96, 3.82]
10	[68, 127, 110]	[61.55, 139.42, 115.67]	[6.45, 12.42, 5.67]
11	[68, 127, 110]	[56.65, 111.74, 96.72]	[11.35, 15.28, 3.28]

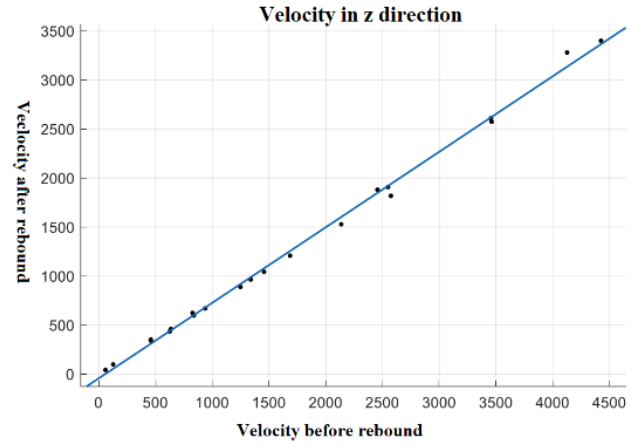


Fig. 5. Approximate  $e_t$ .

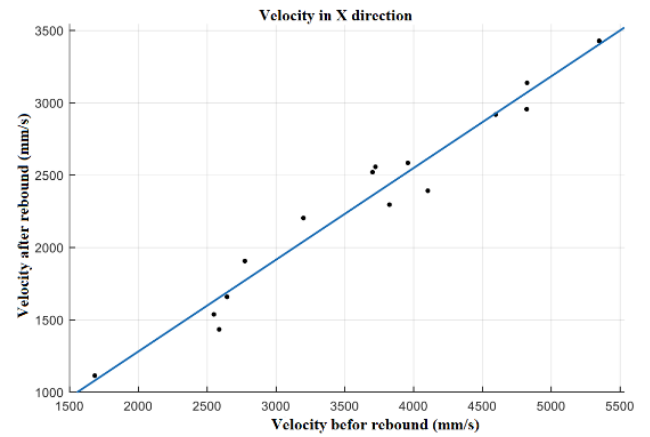


Fig. 6. Approximate  $k_t$ .

2) *Trajectory prediction*: In experiment, six consecutive coordinates are collected and the least squares method is used to determine the initial states. The result was given in (26):

$$\begin{cases} x = -2706t^2 + 6003t + 329.6 \\ y = 368.4t^2 + 45.34t + 258.9 \\ z = -20530t^2 + 1866t + 466.5 \end{cases} \quad (26)$$

where (x, y, z) are the positions of the ball in millimeters, and t is the time in milliseconds. Based on the approximate equations, the initial states of the ball are determined:

$$\begin{cases} x_0 = 329.6mm, v_{x0} = 6003mm/s \\ y_0 = 258.9mm, v_{y0} = 45.34mm/s \\ z_0 = 466.5mm, v_{z0} = 1866mm/s \end{cases} \quad (27)$$

Finally, the trajectory of the ball was predicted according to the equations (21) and (23). The results are shown in Fig 7. Table IV lists the prediction results for ten landing points and errors.

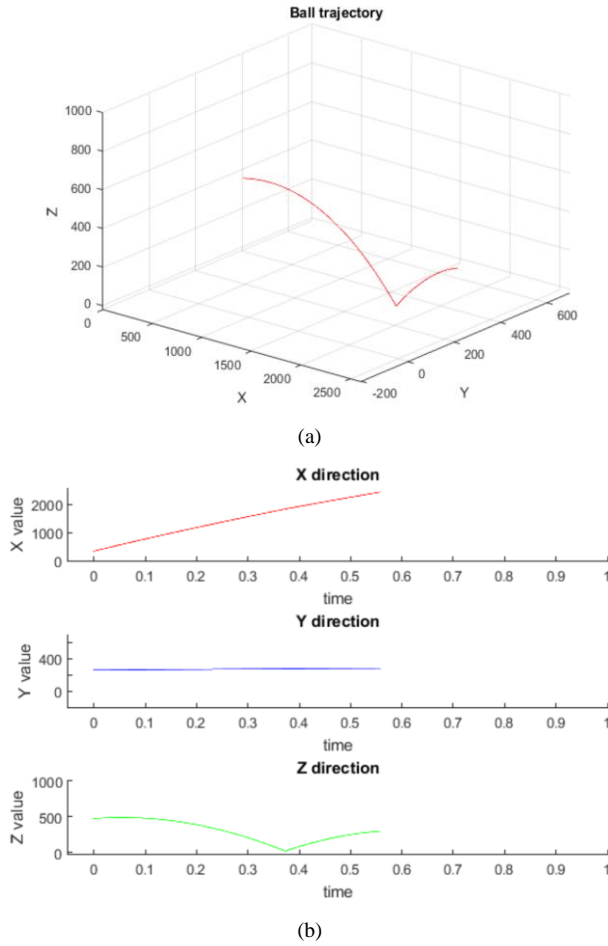


Fig. 7. Trajectory prediction experiment. (a) 3D trajectory (b) Trajectory in X, Y and Z direction.

TABLE IV. LANDING POINTS AND ERRORS.

Index	Actual (x, y) mm	Predict (x, y) mm	Error (x, y) mm
1	[1790, 330]	[1794.09, 259.32]	[4.09, 70.68]
2	[1660, 305]	[1602.10, 300.40]	[2.10, 4.6]
3	[1882, 358]	[1722.24, 308.94]	[79.76, 49.09]
4	[1798, 398]	[1705.57, 339.26]	[92.43, 58.74]
5	[1852, 345]	[1736.29, 295.24]	[115.71, 49.76]
6	[1884, 285]	[1713.27, 225.13]	[170.73, 59.87]
7	[2157, 235]	[2057.86, 215.38]	[99.14, 19.62]
8	[1725, 456]	[1694.35, 448.57]	[30.65, 7.43]
9	[1857, 373]	[1902.87, 313.28]	[45.87, 59.72]
10	[1992, 390]	[1884.36, 426.19]	[107.64, 36.19]

## V. CONCLUSION

This paper has developed the vision system for the ping-pong robot system. The robust image processing algorithm is based on the multi-threshold segmentation with the

combination of HSV threshold, Subtract Background method and the shape features threshold. The 3D position of the ball is computed using triangular method. The aerodynamic model and the rebound model are established to predict the trajectory of the ball. Experiments show the algorithm can detect the ball being affected by noise object, the 3D measurement have satisfied accuracy and the trajectory of the ball can be predicted using models which is established in the research.

## REFERENCES

- [1] R. L. Andersson, *A Robot Ping-Pong Player: Experiment in Real-Time Intelligent Control*. Cambridge, MA: MIT Press, 1988.
- [2] Z. Zhang, D. Xu, and J. Yu, "History and latest development of robot ping-pong player," in *Proc. 7th World Congr. Intell. Control Autom.*, Chongqing, China, pp. 4881–4886, Jun. 25–27, 2008.
- [3] M. Matsushima, T. Hashimoto, M. Takeuchi, and F. Miyazaki, "A learning approach to robotic table tennis," *IEEE Trans. Robot. Autom.*, vol. 21, no. 4, pp. 767–771, Aug. 2005.
- [4] R. Sabzevari, S. Masoumzadeh, and M. R. Ghahroudi, "Employing ANFIS for object detection in robo-pong," in *Proc. Int. Conf. Artif. Intell.*, Jul. 14–17, 2008, pp. 707–712.
- [5] H. Fassler, H.A. Vasteras, and J.W. Zurich, "A robot ping pong player: optimized mechanics, high performance 3d vision, and intelligent sensor control," *Robotersysteme*, 1990.
- [6] Acosta, L. and Rodrigo, J.J. and Mendez, J.A. and Marichal, G.N. and Sigut, M., "Ping-pong player prototype", *IEEE Robotics and Automation Magazine*, vol. 10, 2003, pp 44-52.
- [7] Modi. KP, Sahin, F, Saber, E," An Application of Human Robot Interaction: Development of a Ping-Pong Playing Robotic Arm". In: *Systems, Man and Cybernetics, 2005 IEEE International Conference on*. 10-12 Oct 2005. IEEE. vol. 2, pp. 1831–1836.
- [8] Zhang, Z. and Xu, D. and Tan, M., "Visual Measurement and Prediction of Ball Trajectory for Table Tennis Robot", *Instrumentation and Measurement*, *IEEE Transactions on*, vol. 59, 2010, pp 3195-3205.
- [9] Li, H., Wu, H., Lou, L., Kühnlenz, K., & Ravn, O., "Ping-Pong Robotics with High-Speed Vision System," *12th International Conference on Control Automation Robotics & Vision (ICARCV)*, 2012, pp. 106 - 111.
- [10] Sabzevari, R. and Shahri, A. and Fasih, A. R. and Masoumzadeh, S. and Ghahroudi, M.R., "Object detection and localization system based on neural networks for Robo-Pong," *Proceedings of the 5th International Symposium on Mechatronics and Its Applications*, 2008, pp 1-6.
- [11] Christoph H. Lampert, Jan Peters, "Real-Time Detection of Colored Objects In Multiple Camera Streams With Off-the-Shelf Hardware Components," *Journal of Real-Time Image Processing*, vol. 7, 2010, pp 31-41.
- [12] Xiaopeng Chen and Zhangguo Yu, "Dynamic model based ball trajectory prediction for a robot ping-pong player," *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pp. 603-608, 2010.
- [13] Han Bao, Xiaopeng Chen, ZhanTao Wang, Min Pan, Fei Meng, "Bouncing Model for the Table Tennis Trajectory Prediction and the Strategy of Hitting the Ball," *Proceedings of 2012 IEEE International Conference on Mechatronics and Automation August 5 - 8, Chengdu, China*, pp 2002-2006, 2012.

# Effect of the Welding Parameters on Mechanical Properties of AA5083 Friction Stir Welding

Phan Thanh Nhan  
Engineering Mechanical Faculty  
HCMC University of Technology and Education  
Ho Chi Minh City - Vietnam  
nhanpt@hcmute.edu.vn

**Abstract**— With high strength and good corrosion resistance, aluminum alloy 5083 is considered a key material in the manufacture of high speed craft. Welding 5083 aluminum alloy by melting welding methods often significantly reduces structural life. Friction welding is considered an effective technological solution for welding difficult alloys and 5083 aluminum alloys in particular. In this study, the stir friction welding of AA5083 aluminum alloy plate with the thickness of 3 mm was fabricated and tested the effect of welding parameters: welding regime  $\omega$  (rev/min) and transverse speed of weld tool  $v$  (mm/min) on the microstructure and mechanical properties of the weld. Generally, the weld can be achieved very good quality, with no significant defects except for "kissing bonds defect". The grain structure at welding area varies significantly. The grain structure at the weld center is much finer than the base metal. The hardness in the welding area is significantly reduced compared to the base material. Experimental result shows that the tensile strength of the weld increases with increasing welding speed. The weld reaches a tensile strength of over 85% (compared to the base metal) in the welding regimes of 1.40, 1.75 and 3.33 rev/mm.

**Keywords**— Stir friction welding, 5083 aluminum alloy, weld structure, hardness, mechanical properties.

## I. INTRODUCTION

Nowadays, aluminum alloy is a material used in many different applications with outstanding advantages such as light, high strength, corrosion resistance... Applications can be mentioned as shipbuilding, car part manufacturing, more specifically used in the aircraft bodies making. There are instances where different series of aluminum alloys are to be welded due to its requirement in the varied service conditions [1]. Although there are many dominances, the biggest drawbacks of aluminum alloys is poor welding. Friction stir welding (FSW) is considered an effective technology solution to solve this problem, this is a welding method invented by the British Academy (TWI) in 1991 [2]. Welding friction stirring is an advanced welding method with the basic principle of using a rotating tool with a material which has melting point higher than welding materials and plastic deformation of welding materials generated by friction. The plastic deformation of the welding material stirred with the rotating circular motion will form the bonds (Fig. 1) [3]. In many studies, it has been demonstrated that the application of FSW for aluminum alloy has better weld quality than conventional methods [4,5]. In order to have a good mechanical properties of weld by FSW, depending on the types of aluminum alloys, it is necessary to select the appropriate welding parameters.

This paper will focus on the effect of two welding parameters: welding regime ( $\omega$  – rev/min) and transverse speed of weld tool ( $v$  – mm/min) on tensile strength of the AA5083 aluminum alloy weld.

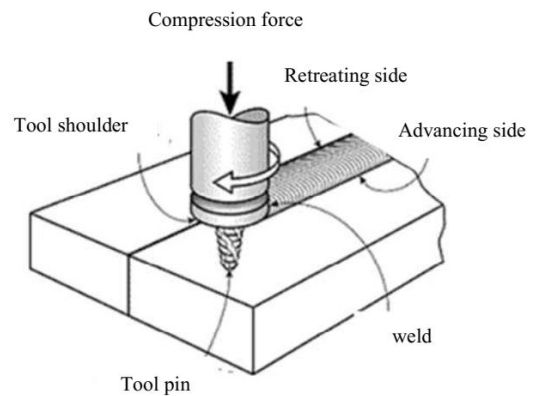


Fig. 1. Friction stir welding process [3]

## II. MATERIALS AND EXPERIMENTS

Butt joint by FSW of two 3 mm thick aluminum alloy plates AA5083 are made from friction stir welding machine. The welding tool has a shoulder diameter of 16 mm, pin diameter of 4 mm and pin length of 2.8 mm (cylindrical pin). The material selected for making welding tools is H13. After designing and manufacturing, welding tool are heat-treated and tested to meet the desired hardness. Two 5083 aluminum alloy plates are fixed on the table by jigs (fig. 2a). The table is angled at 2 ° to the direction of the tool pin (fig. 2b). The welding process is shown in fig. 2c and the completed weld is displayed in fig. 2d.

After the weld is completed in many different regimes, the weld is screened for weld quality by impregnating with solution of 2 ml HF, 3 ml HCl, 20 ml HNO<sub>3</sub>, 175 ml H<sub>2</sub>O [6] and is observed with a 1000x magnification camera (fig. 3a). The microstructure of the weld is carried out by electrochemical corrosion method on ElectroMet 4 (fig. 3b) and is observed by Olympus CK40M microscope. The hardness of the weld is measured on Rockwell machine with HRB scale, using a ball bearing with a load of 100 kg (fig. 3c). Tensile strength of the weld is tested on Instron 3366 machine with a maximum tensile force of 10 kN, speed of 5 mm/min (fig. 3d). The sample is fabricated according to ASTM E290 [7] (fig. 4).



### III. RESULTS AND DISCUSSION

#### A. The microstructure of the weld

After the weld is completed in many different regimes, the weld is screened at the cross section in order to assess the weld's quality. At regime of  $\omega/v = 2.33$  rev/mm (fig. 5b) cracks appeared at the bottom of the weld. At regime of  $\omega/v = 1.75$  rev/mm (fig. 5c) appeared kissing bonds defect [8]. However, in another study it was shown that kissing bonds defect is not dependent on welding parameters and is difficult to remove. When testing the tensile strength of the weld, this defect did not significantly affect the tensile strength. At regime of  $\omega/v = 1.40$  rev/mm, no defect displayed (fig. 5d).

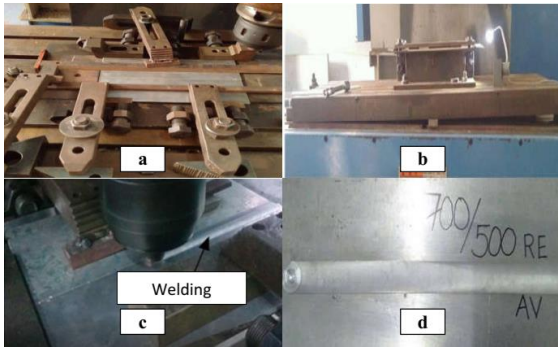


Fig. 2. The process of installing and fabricating welds

- 2a. Jigs                      2b. The angled set table  
2c. The welding process    2d. The completed weld



Fig. 3. Experimental process

- 3a. Camera                      3b. ElectroMet 4  
3c. Rockwell machine        3d. Instron 3366

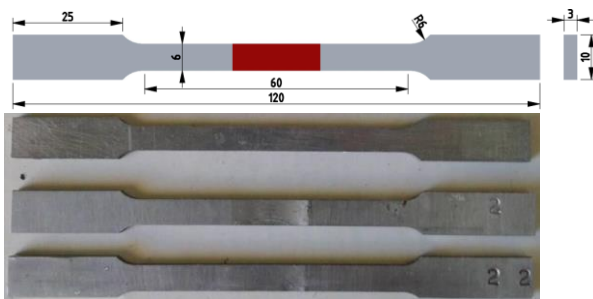


Fig. 4. Tensile samples (ASTM E290)

Fig. 6 shows the weld microstructure at  $\omega/v = 1.40$  rev/mm. It clearly showed the different grain structure in the zones of the weld. (I) (Base Metal – BM) has a large grain size due to unaffected temperature, about  $40 \div 60 \mu\text{m}$ . (II) (Heat

Affect Zone -HAZ) is the zone close shoulder of welding tool, which is affected by temperature due to friction but not plastic deformation. The grain structure in this zone has almost no change compared to the base material zone, grain size in the range of  $33 \div 50 \mu\text{m}$ . (III) (Thermo Mechanically Affect Zone – TMAZ) is the zone under the shoulder of welding tool and develop plastic deformation by friction heat. At this zone, there is a clear change in grain structure compared to HAZ, grain size in the range of  $26 \div 37 \mu\text{m}$ . (IV) (Stir Zone – SZ), where the material has the most plastic deformation, the grain size also changes most clearly, grain size in the range of  $16 \div 24 \mu\text{m}$ .

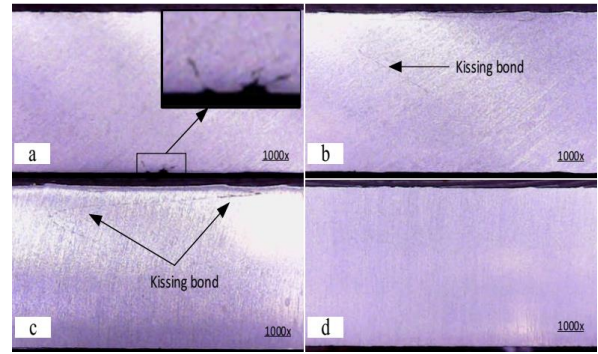


Fig. 5. Types of welding defects

- 5a.  $\omega/v = 3.50$  rev/mm    5b.  $\omega/v = 2.33$  rev/mm  
5c.  $\omega/v = 1.75$  rev/mm    5d.  $\omega/v = 1.40$  rev/mm

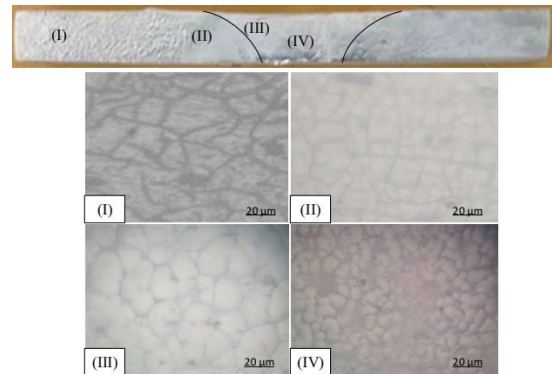


Fig. 6. The weld microstructure at  $\omega/v = 1.4$  rev/mm.

#### B. Distribution of weld hardness

The hardness of the weld is measured at a cross-section with the HRB scale. The results show that all welds have reduced hardness from both sides RS and AS to weld center (fig. 7). At the center of the weld (SZ) there has been the smallest hardness. The higher the welding speed is, the lower the hardness at the center of the weld is. At the regime of  $\omega/v = 1.40$  rev/mm the hardness has lowest value. The hardness distribution clearly shows the influence of welding parameters on the mechanical properties of the weld. This change is due to the influence of temperature leading to a change in the microstructure of the weld.

#### C. Tensile strength of welds

Results of the tensile strength of welds in different regimes are shown in Table 1. Evaluation of weld quality is carried out by comparison with the tensile strength of the base material.

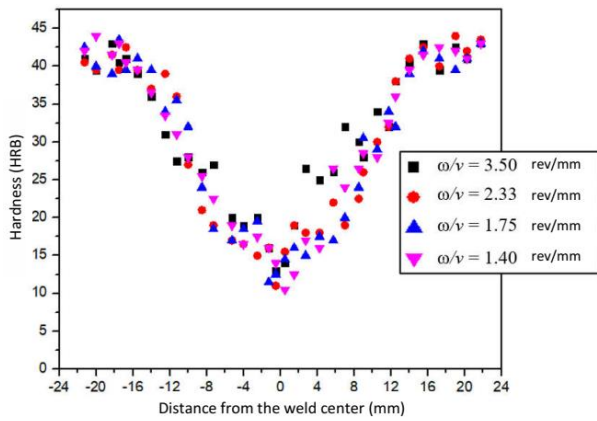


Fig. 7. Weld hardness distribution at different regimes.

TABLE I. RESULTS OF TENSILE TESTS

Regimes $\omega/v$ (rev/mm)	Sam- ple	Tensile strength $\sigma_{max}$ (MPa)	Mean value of tensile strength (MPa)	Deforma- tion at $\sigma_{max}$ (%)	Mean value of defor- mation (MPa)
Base Material	1	314.12	313.15	15.30	14.26
	2	312.18		13.22	
3.50	1	209.79	211.22	3.06	3.22
	2	218.82		3.44	
	3	204.99		3.16	
2.33	1	277.41	264.73	8.15	7.57
	2	257.39		6.38	
	3	259.19		8.19	
1.75	1	278.98	269.19	8.30	7.54
	2	270.18		7.06	
	3	258.43		7.26	
1.40	1	278.01	270.97	7.16	7.49
	2	260.52		7.34	
	3	274.39		7.98	

Experimental results show that all welds are destroyed in the welding zone. At the regime of  $\omega/v = 3.50$  rev/mm and of  $\omega/v = 1.40$  rev/mm the fail section appears at the center of the weld. At the regime of  $\omega/v = 2.33$  rev/mm and of  $\omega/v = 1.75$  rev/mm, destruction position is outside the weld center and on the AS. Fig. 8 shows cracks when samples are destroyed at cross-sections of welds. At the regime of  $\omega/v = 2.33$  rev/mm and of  $\omega/v = 1.75$  rev/mm, destruction cracks are formed along the kissing bonds line; however, the tensile strength in these two regimes is still very good. This shows that kissing bond lines do not significantly affect the tensile strength of the weld. At the regime of  $\omega/v = 1.40$  rev/mm, destructive cracks have a 45° oblique profile similar to the crack of the base material, so in this regime the tensile strength of the weld is best.

Tensile strength and deformation of welds decrease as welding speed increases. At the regime of  $\omega/v = 3.50$  rev/mm, the weld has the lowest tensile strength and deformation. The reason is that when the welding speed is low, the friction time between the tool and the welding material is longer and then the friction heat generated is very high, making the weld brittle. At the regimes of  $\omega/v = 2.33$ , 1.75 and 1.40 rev/mm, tensile strength and deformation increase significantly. At the regimes of  $\omega/v = 1.40$  rev/mm, the weld has the highest tensile strength of 270.97 MPa reach 86.53% compared to the tensile strength of base material. From experimental results, at the

regimes of  $\omega/v = 2.33$ , 1.75 and 1.40 rev/mm, the temperature produced by friction is consistent with the formation and change of the microstructure of the weld and hence the tensile strength of the weld achieves good quality. The influence of welding parameters on the tensile strength and deformation of welds is shown in fig. 9.

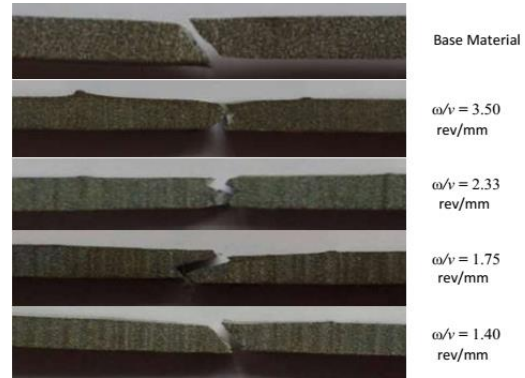


Fig. 8. Failure track at cross sections of weld in welding regimes.

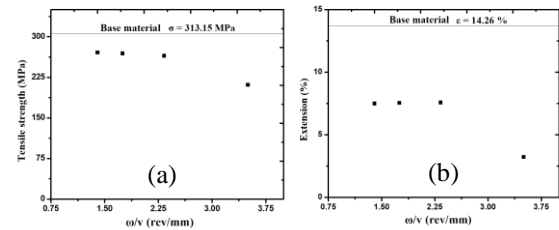


Fig. 9. Influence of welding regimes on

- (a) Tensile strength of the weld  
(b) Deformation of the weld

#### IV. CONCLUSION

Stirring friction weld for aluminum alloy AA5083 (3mm thick plates) has been fabricated successfully. The effect of two parameters of spin speed  $\omega$  (rev/min) and transverse speed of weld tool  $v$  (mm/min) on the microstructure, hardness and tensile strength of the weld was investigated. Through the experiment results, it was found that the welding parameters are suitable and ensure the weld quality with negligible defects and high mechanical properties of the weld.

At the regimes of  $\omega/v = 2.33$  rev/mm and  $\omega/v = 1.75$  rev/mm, although the structure of weld still exists with kissing bonds, it does not significantly affect weld quality. At the regime of  $\omega/v = 1.40$  rev/mm, the weld completely remove defects. The weld has the lowest hardness at the center of the weld (SZ). Tensile strength and deformation are directly proportional to the transverse speed of weld tool. Welds have good tensile strength and deformation in welding regimes of 2.33, 1.75 and 1.40 rev/mm. The highest tensile strength of the weld reaches 86.53% compared to the base material.

#### REFERENCES

- [1] RajKumar. V, VenkateshKannan. M Sadeesh. P, Arivazhagan. N, Devendranath Ramkumar. K, "Studies on effect of tool design and welding parameters on the friction stir welding of dissimilar aluminium alloys AA 5052 - AA 6061", Procedia Engineering, 75, 2014, pp. 93-97
- [2] Rajiv Sharan Mishra, Partha Sarathi De Nilesh Kumar, "Friction Stir Welding and Processing", Science and Engineering, 2014, pp. 3

- [3] W M Thomas, K I Johnson, and C S Wiesner, "Friction stir welding - recent developments in tool and process technologies", *Advanced Engineering Materials*, Volume 5, Issue 7, July 2003, pp. 485-490.
- [4] A. Sik, M. Onder, "Comparison between mechanical properties and joint performance of AA2024-O aluminium alloy welded by friction stir welding and TIG processes", *ResearchGate*, 50, 2012, pp. 131-137.
- [5] Haitham Kassim Mohammed, "A comparative study between friction stir welding and metal inert gas welding of 2024-T4 Aluminum alloy", *ARPJ Journal of Engineering and Applied Sciences*, Vol. 6, No. 11, 2011, pp. 36-40.
- [6] Mostafa M. El-Sayed, Ahmed Y. Shash, Tamer S. Mahmoud and Mahmoud Abd Rabbou, "Effect of Friction Stir Welding Parameters on the Peak Temperature and the Mechanical Properties of Aluminum Alloy 5083-O". *Advanced Structured Materials*, 2018, pp. 11-25.
- [7] Standards ASTM, E08/E8M – 13a, "Test Methods for Tension Testing of Metallic Materials", 2013.
- [8] Paul Kah\*, Richard Rajan, Jukka Martikainen and Raimo Suoranta, "Investigation of weld defects in friction-stir welding and fusion welding of aluminium alloys", *International Journal of Mechanical and Materials Engineering*, 2015, pp. 1-10.
- [9] Duong Dinh Hao, Masakazu Okazaki, Tran Hung Tra, Quach Hoai Nam, "Defects Morphology in the Dissimilar Friction Stir Welded T-lap Joints of AA7075 and AA5083", *Advances in Engineering Research and Application*, 2018, pp. 210-216.



# An improvement of Disk Aware Discord Discovery Algorithm for Discovering Time Series Discord

Nguyen Thanh Son

Faculty of Information Technology  
Ho Chi Minh City University of  
Technology and Education, Vietnam  
sonnt@fit.hcmute.edu.vn

**Abstract**— A time series is a sequence of data points where each point represents a value at a given point in time. The problem of discovering discord in time series has received a lot of attention lately. Time series discord is a subsequence of a long time series which is the most different from all the rest of the time series subsequences. In this work, we propose an improvement of Disk Aware Discord Discovery (DADD) algorithm for time series discord discovery. The improvement is based on symbolic aggregate approximation method associated with a hash bucket structure to speed up the selection for discord candidates. The experimental results showed that our improved algorithm outperforms the original method, Disk Aware Discord Discovery, in terms of runtime while the accuracy is the same.

**Keywords**— Time Series, Time Series Discord, Symbolic Aggregate Approximation, Discord Discovery

## I. INTRODUCTION

A time series is a series of real numbers which represent data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Time series data are used in a lot of application areas. Time series discord discovery is one of problems which are interested in research in recent years.

Time series discord is defined as a subsequence which is maximally different to all the rest of subsequences of a longer time series. Time series discord discovering has been used for fault diagnostics, intrusion detection, data cleansing and so on.

Many algorithms for discovering time series discord has been proposed since a formal definition of time series discord introduced in 2005 by Keogh et al. [1]. Most of them usually assume that the data reside in main memory and resort to multiple scans of the databases to discover a discord. For many real-world problems this is not be the case. So, Yankov et al., proposed a new algorithm in which time series discord can be discovered with only two linear scans of the disk and a tiny buffer of main memory. Their proposed algorithm is exact and it is very simple to implement [2].

In our work, we introduce an algorithm which is an improvement of Disk Aware Discord Discovery algorithm proposed by Yankov et al. The improvement is made in the discord candidate selection phase and based on two things: (1) using well-defined and well-documented dimensionality reduction power of PAA and the reduction is automatically carried over to the SAX representation and (2) using a hash basket structure to select discord candidates which is based on comparing symbol sequences. So, we can speed up the selection for discord candidates.

We experimented the proposed algorithm on time series datasets of various areas. The experimental results show that our improved algorithm outperforms the original method, Disk Aware Discord Discovery, in terms of runtime while the accuracy is the same.

The rest of the paper is organized as follows. In Section 2 we review related work and basic concepts. Section 3 describes our approach for discovering time series discord. Section 4 presents our experimental evaluation on different datasets. In section 5 we include some conclusions and suggestions for future work.

## II. RELATED WORKS AND BASIC CONCEPTS

### A. Related Works

Many algorithms have been introduced to solve the time series discord discovery problem since it was formalized in 2005 [1]. In [1] Keogh et al. proposed a fast heuristic technique (called Hot SAX) for pruning quickly the data space and focusing only on the potential discords. Fu et al. proposed a new algorithm based on Haar Wavelet transform to determine dynamically the word size for the compression of subsequences [3]. In [4] Bu et al. proposed a new method called WAT (Wavelet and augmented TRIE) which is based on Haar Wavelet transform and augmented TRIE to mine the top-k discords from time series data. Chuah et al. proposed a new anomaly detection method for discovering discord in ECG dataset [5]. It is based on time series analysis in order to determine whether a stream of real-time sensor data contains any abnormal heartbeats. If anomaly exists, that time series segment will be transmitted via the network to a physician so that experts can further diagnose the problem and take appropriate actions.

In [6] Lin et al. introduced a new approach for the anomaly detection problem. First, this method uses subseries join to obtain the similarity relationships among subseries of the time series data. Then it converts the anomaly problem to graph-theoretic problem which can be solve by existing graph-theoretic algorithm. A new method for time series discord discovery, called HOTiSAX, proposed by Buu et al. [7]. This algorithm incorporates iSAX (indexable Symbolic Aggregate approXimation) representation in Hot SAX instead of SAX representation. In [8] Khanh et al. proposed a new method for discord discovery in time series, called WATiSAX. This algorithm employs iSAX representation in WAT algorithm to detect discord in time series.

A new method proposed by Luo et al. which exploits a recurrence structure of time series and uses a reference function that makes the search algorithm more efficient and

robust [9]. Jones et al. introduced a new algorithm for discovering anomalies in real valued multidimensional time series [10]. First this method uses an exemplar-based model for detecting anomalies in single dimensional time series, then uses a function that predicts one dimension from a related one. A new algorithm which uses grammar induction to aid anomaly detection without any prior knowledge proposed by Pavel Senin et al. [11]. First, this algorithm discretizes continuous time series values into symbolic form, then it infers a context free grammar. Finally, the algorithm uses its hierarchical structure to effectively and efficiently discover anomalies. In [12] Nguyen T. S. proposed a new algorithm for discovering time series discord based on R\*-tree. This method needs a single scan over the entire time series database and a few times to read the original disk data in order to validate the results. A method for discovering discord in streaming time series data proposed by Chau et al. [13]. This method uses clustering algorithm instead of augmented trie in Hot SAX method.

To discover discords in massive datasets, in [2] Yankov et al. proposed a new method, called Disk aware discord discovery. This method includes two phases: (1) a candidate selection phase and (2) a discord refinement phase. In phase 1, the algorithm performs a linear scan through the database  $T$ . Each  $T_i \in T$  is validated to see if it is likely to be a discord or it is omitted. Phase 2 accepts as an input a candidate set  $C \subset T$ , which is a result from phase 1. It will prune the set  $C$  to retain only the true discord.

### B. Basic Concepts

In this subsection we give the definitions of the terms formally.

**Definition 1. Euclidean distance:** Given two time series  $Q = \{q_1, \dots, q_n\}$  and  $C = \{c_1, \dots, c_n\}$ , the Euclidean distance between  $Q$  and  $C$  is defined as:

$$D(Q, C) = \sqrt{\sum_{i=1}^n (q_i - c_i)^2} \quad (1)$$

The Euclidean distance metric is the simplest method to measure the similarity of time series and has been widely used for pattern matching [14]

**Definition 2. Time series:** A time series is a real value sequence of length  $n$  over time, i.e. if  $T$  is a time series then  $T = (t_1, \dots, t_n)$  where  $t_i$  is a real number.

**Definition 3. Subsequence:** Given a time series  $T = (t_1, \dots, t_n)$ , a subsequence of length  $m < n$  of  $T$  is a sequence  $S = (t_i, \dots, t_{i+m-1})$  with  $1 \leq i \leq n - m + 1$ .

Since all subsequences may potentially be discords, we have to compare any subsequence to all remaining subsequences. However, the best matches of a subsequence tend to be located some points to the left or to the right of the subsequence in question. Such matches are called trivial matches and they have to be excluded from the result of discovering discords.

**Definition 4. Non-trivial match:** Given a time series  $T$ , containing a subsequence  $C_p$  of length  $m$  beginning at position  $p$  and a matching subsequence  $C_q$  beginning at  $q$ , we say that  $C_q$  is a non-trivial match to  $C_p$  if  $|p - q| \geq m$ .

**Definition 5. Time series discord:** Given a time series  $T$ , the subsequence  $C$  of length  $n$  is the most significant discord in  $T$

if the distance to its nearest non-trivial match  $Q$  is largest. It means that for an arbitrary subsequence  $M \in T$ ,  $\min(D(C, Q)) \geq \min(D(M, P))$ , where  $Q, P$  are subsequences in  $T$  and  $Q, P$  are non-trivial matches of  $C$  and  $M$ , correspondingly.

**Definition 6.  $K^{\text{th}}$  Time series discord:** Given a time series  $T$ , the subsequence  $C$  of length  $n$  beginning at position  $p$  is the  $K^{\text{th}}$  significant discord in  $T$  if  $C$  has the  $K^{\text{th}}$  largest distance to its nearest non-trivial match and there is no overlap region between  $C$  and the  $i^{\text{th}}$  discord beginning at position  $q$ , for all  $1 \leq i < K$ . This means  $|p - q| \geq n$ .

## III. OUR APPROACH

All subsequences extracted from a time series  $T$  can form a *subsequence database* in which each subsequence can be regarded as a time series. So, the discussion is limited to the case where the time series database  $T$  contains  $|T|$  separate time series of length  $n$ . If instead the database contains subsequences from a long time series the basic algorithm is unchanged but the trivial matches must be rejected.

The basic intuition behind our proposed algorithm is as follows: in the candidate selection phase we first transform each normalized time series into the Piecewise Aggregate Approximation (PAA) representation and then symbolize the PAA representations into the discrete strings using Symbolic Aggregate approXimation (SAX) method. After that the SAX sequences are divided into baskets in which the same sequences are placed in the same basket. Finally, selected candidates are sequences in baskets with the least number of items. In phase 2, the true discord will be retrieved based on the candidate set by finding on this small candidate set in original space.

Note that, each time series need to normalize to have mean of zero and a standard deviation of one before transforming it to PAA representation because it is understood that it is meaningless to compare time series with different offsets and amplitudes [14].

### A. The zero mean normalization

Given a time series  $T = \{t_1, t_2, \dots, t_n\}$ .  $T$  can be transformed to a normalized sequence  $T' = \{t'_1, t'_2, \dots, t'_n\}$  to have mean of zero and a standard deviation of one by following formular:

$$t'_i = \frac{(t_i - \text{mean}(T))}{\text{std}(T)} \quad (2)$$

where,  $\text{mean}(T)$  is the mean value and  $\text{std}(T)$  is the standard deviation of time series  $T$ .

### B. The PAA representation

Given a time series  $T$  of length  $n$ ,  $T = (t_1, t_2, \dots, t_n)$ . To transform this time series into the  $w$  dimensional space ( $w \ll n$ ) using PAA method we can do as follows: first, the time series  $T$  is divided into  $w$  equal sized segments, then the mean value of the data within each segment is calculated. The vector of these values is an approximation representation of time series  $T$ . It means that a time series  $T$  of length  $n$  can be represented approximately in a  $w$  dimensional space by a vector  $V = v_1, v_2, \dots, v_w$ , where  $v_i$  is calculated by the following formular [15]:

$$v_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} t_j \quad (3)$$

### C. The SAX discretizing method

Given the normalized time series which have highly normal distribution. The breakpoints that will produce equal-sized areas under Gaussian curve can be simply determined by looking them up in a statistical table. For example, in table 1 we show the breakpoints that divide a normal distribution of equal-sized regions from 3 to 8.

The breakpoints here are sorted list of numbers  $\beta = \beta_1, \dots, \beta_{a-1}$  such that the area under a Gaussian curve from  $\beta_i$  to  $\beta_{i+1} = 1/a$ , where  $a$  is the number of equal-sized areas under Gaussian curve ( $\beta_0 = -\infty$  and  $\beta_a = \infty$ )

TABLE I. A LOOKUP TABLE FOR DETERMINING BREAKPOINTS WITH A FROM 3 TO 8.

$\beta_i \backslash a$	3	4	5	6	7	8
$\beta_1$	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15
$\beta_2$	0.43	0	-0.25	-0.43	-0.57	-0.67
$\beta_3$		0.67	0.25	0	-0.18	-0.32
$\beta_4$			0.84	0.43	0.18	0
$\beta_5$				0.97	0.57	0.32
$\beta_6$					1.07	0.67
$\beta_7$						1.15

Once the breakpoints have been determined, a PAA representation of time series can be discretized as follows: all PAA coefficients that are below the smallest breakpoint are mapped to the symbol 'a', all coefficients greater than or equal to the smallest breakpoint and less than the second smallest breakpoint are mapped to the symbol 'b', and so on. Figure 1 illustrates an example of a PAA representation of time series is mapped into SAX symbols with  $a = 3$  (using three symbols 'a', 'b' and 'c'). In this example, the PAA representation is mapped into SAX sequence 'baabcbbc'.

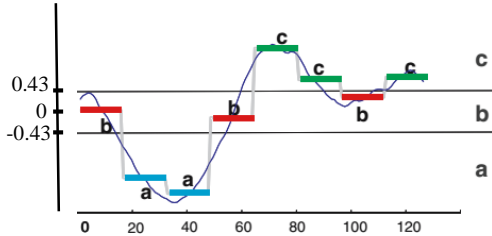


Fig. 1. An example of a PAA representation of time series is mapped into SAX symbols.

Note that, assuming normal distribution is really true for a large group of the time series except a small subset of time series does not follow. So, the efficiency can be slightly deteriorated if time series is not obeyed normal distribution. However, the correctness of the algorithm is not affected. The correctness of the algorithm is ensured by the lower-bounding property of the distance measure in the SAX space [15].

### D. The discord discovering method

So far we have presented some related fundamental knowledge which is the focus of our work. Here we are going to propose an efficient algorithm for discovering the top discord in a time series database.

The discord discovering algorithm includes two phases: (1) a candidate selection phase and (2) a refinement phase.

In phase 1, the algorithm will find all discord candidates based on comparing SAX sequences. To do this the algorithm need to performs a scan through the database. For each time series which are normalized to have mean of zero and a

standard deviation of one, they are transformed to SAX sequences based on PAA dimensional reduction method. Then these SAX sequences are hashed into baskets such that each basket only contains the same sequences. Finally, the selected discord candidates are sequences in the baskets with the least number of items.

Figure 2 illustrates the algorithm for phase 1. In figure 2,  $S$  is a list of SAX representations (line 1) and  $L$  is a list of baskets (line 2). The algorithm check if a SAX sequence is the same as one of the items in  $L$ . If the comparison is true, the sequence is going to add to the corresponding basket and the algorithm will exit the loop (lines from 7 to 12). In case the sequence is not the same as any item in  $L$ , the algorithm is going to put it in a new basket (line 13). At the end of algorithm, the basket with the least number of items are selected as discord candidates (lines from 15 to 17).

#### Algorithm Selecting discord candidates.

**Input:**  $T$ : normalized time series database  
 $w$ : dimensional number in feature space  
 $a$ : number of symbols using in SAX representation  
**Output:**  $C$ : list of discord candidates  
1:  $S[1] = \text{SAX\_representation}(T_1, w, a)$   
2:  $L[1] = \{S[1]\}$   
3:  $C = \text{null}$   
4: **For**  $i = 2$  to  $|T|$  **do** {  
5:  $S[i] = \text{SAX\_representation}(T_i, w, a)$   
6:  $\text{found} = \text{false}$   
7: **While**  $(\forall L[j] \in L \text{ or } \text{found} = \text{true})$  **do** {  
8: **If**  $(S[i] == L[j][1])$  **then** {  
9:  $L[j] = L[j] \cup i$   
10:  $\text{found} = \text{true}$   
11: }  
12: }  
13: **If**  $(\text{!found})$  **then** Add  $i$  to a new basket in  $L$   
14: }  
15: **For**  $\forall i \in L$  **do**  
16: **if**  $(L[i]$  has the least number of items) **then**  
17:  $C = C \cup L[i]$

Fig. 2. The algorithm for discovering discord candidates based on SAX representation.

The refinement phase in our work is similar to phase 2 proposed in [2]. In this phase, the algorithm receives the discord candidate set  $C$ , the result in phase 1, as an input and then prunes the set  $C$  in order to remove false discords. Figure 3 illustrates the algorithm for this phase which is the same as the discord refinement algorithm proposed in [2]. In this algorithm,  $C$  is a list of the discord candidates discovered in phase 1 but it is unknown which items in  $C$  are the true discords and what their actual discord distances are. The algorithm needs a scan through the time series database. For each sequence in the database, it is compared to the discord candidates. The actual distance between a candidate and a sequence in the database is calculated by using Euclidean distance associated with the idea of early abandoning for optimization (line 5). The idea of early abandoning is performed as follows: when the Euclidean distance is calculated for a pair of time series, if the cumulative sum is greater than the current best-so-far distance at a certain point we can abandon the calculation because this pair of time series is not a best match [16]. Figure 4 shows the intuition behind this technique. In this example the current best-so-far distance is supposed of 11. At the point the squared Euclidean distance of 121 we can stop this calculation.

**Algorithm** Discovering true discord.

---

**Input:**  $T$ : normalized time series database  
 $C$ : set of discord candidates  
 $r$ : discord range threshold

**Output:**  $C$ : list of the true discord  
 $dist$ : list of nearest neighbor distance to the discords

```

1: For  $i = 1$  to  $|C|$  do  $dist_i = \infty$ 
2: For  $\forall T_i \in T$  do {
3:   For  $\forall C_j \in C$  do {
4:     If  $(T_i = C_j)$  then continue
5:      $d = D\_Early\_abandon(T_i, C_j, dist_j)$ 
6:     If  $(d < r)$  then {
7:        $C = C \setminus C_j$ 
8:        $dist = dist \setminus dist_j$ 
9:     }
10:    else {
11:       $dist_j = \min(dist_j, d)$ 
12:    }
13:  }

```

---

Fig. 3. The algorithm for the refinement phase [2].

There are three cases after the true distance between a sequence  $T_i$  in the time series database and a discord candidate  $C_j$  in  $C$  has been calculated [2]:

- In the case that the true distance between  $T_i$  and  $C_j$  is greater than the current value of  $dist_j$  we do nothing because  $T_i$  is not a nearest neighbor of  $C_j$ .
- If the true distance between  $T_i$  and  $C_j$  is less than  $r$ ,  $C_j$  can be removed from the candidate set  $C$  because it can not be a true discord (line 6 and line 7).
- If the true distance between  $T_i$  and  $C_j$  is greater than or equal to  $r$  (but still less than the current value of  $dist_j$ ), the current distance to the nearest neighbor,  $dist_j$ , is updated (line 10).

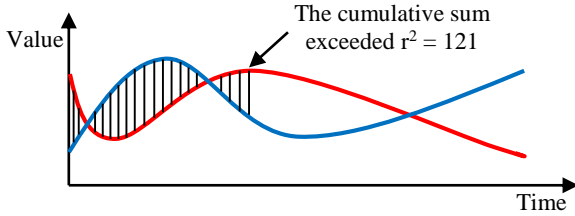


Fig. 4. An illustration of the idea of early abandoning technique.

## IV. EXPERIMENTAL RESULTS

For the experiments, we implemented all the algorithms with Visual Microsoft C# and all the experiments are conducted on a Core i5, 2.4 GHz, 4.0 GB RAM. We tested on three different publicly available datasets: EEG, Consumer and discord anomaly which are published in the internet for free public download. We compare our proposed approach to the original method, Disk Aware Discord Discovery (DADD), in term of run time and accuracy. We conducted the experiments on the datasets with cardinalities ranging from 4000 to 24000 and the different lengths of discord from 64 to 1024. With SAX representation, we set the number of breakpoints to 3. For our proposed method, we use the compression ratio of 1:1 to be fair with the Disk Aware Discord Discovery method.

The accuracy of the proposed discord discovery algorithm is basically based on human analysis of the discords

discovered by that algorithm ([1], [4], [5], [17]). That means if the discords identified by a proposed algorithm on most of the test dataset are almost the same as those identified by the baseline discord discovery algorithm (here the brute-force algorithm is chosen as the baseline algorithm), we can say that the proposed discord discovery algorithm and DADD method bring out the same accuracy as the baseline algorithm.

For brevity, we only report some typical experimental results. Figure 5 shows the running time from the experiments of two methods on the EEG dataset of size 4000 with different discord lengths. Figure 6 displays the running time from the experiments of two methods on the EEG dataset with different sizes. The fixed discord length is 512. Figure 7 lists the running time from the experiments of two methods on the three different datasets with the fixed size of 10000. The fixed discord length is 512.

As we can see in these figures, the running time of our proposed method is less than that of DADD method. They also show that the different in runtime is negligible in the case of short discord length or small database size. But it will be much different when the discord length or database size increases.

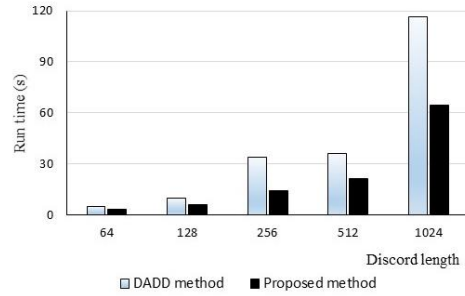


Fig. 5. The running time of two methods on the EEG dataset with different discord lengths.

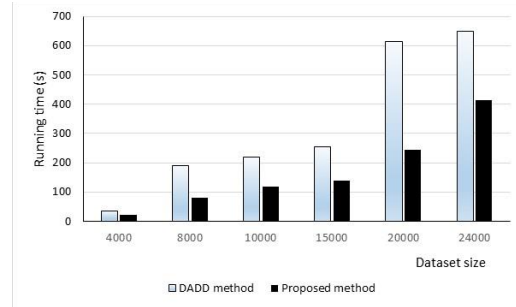


Fig. 6. The running time of two methods on the EEG dataset with different size and fixed discord length of 512.

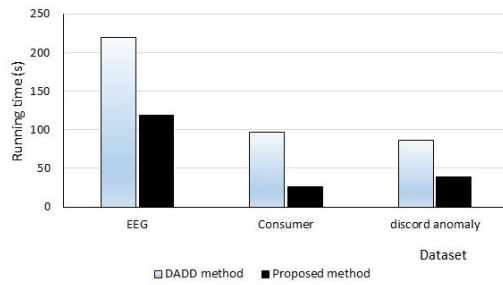


Fig. 7. The running time of two methods on three datasets with fixed size of 10000 and fixed discord length of 512.

Figure 8 shows the plots of EEG dataset (above figure) and discord of length 512 discovered from the experiments of

three methods on EEG dataset of size 10000. As you can see in figure 8, the discords discovered by three methods are exactly the same. Experiment on the remain datasets also shows similar results.

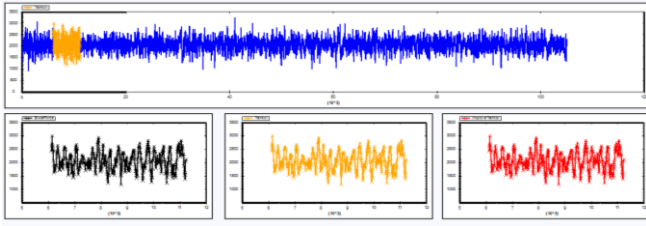


Fig. 8. The experimental result on EEG dataset (above figure) with discord length 512. The discords discovered by brute-force algorithm (below and left figure) and by DADD method (below and center figure) and by proposed method (below and right figure)

## V. CONCLUSION

We introduce a new algorithm to discover discord in a long time series which is an improvement of Disk Aware Discord Discovery algorithm proposed by Yankov et al. This algorithm uses SAX representation associated with a hash basket structure to speed up the selection for discord candidates and the Euclidean distance combined with the idea of early abandoning to speed up the search for matching neighbors of a subsequence. The experiments on the different datasets demonstrate that our proposed method outperforms the original method in terms of runtime while the accuracy is the same.

## REFERENCES

- [1] Keogh, E., Lin, J. and Fu, A., "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence". In: *the 5th IEEE International Conference on Data Mining (ICDM'05)*, pp. 226-233. Houston, TX (2005).
- [2] Yankov, D., Keogh, E. and Rebbapragada, U., "Disk aware discord discovery: Finding unusual time series in terabyte sized datasets". *Knowledge and Information Systems*, vol. 17, no. 2, pp. 241-262 (2008).
- [3] Fu, A., Leung, O., Keogh, E. and Lin, J., "Finding Time Series Discords Based on Haar Transform". In: *Lecture Notes in Computer Science, Advanced Data Mining and Applications*, vol. 4093, pp. 31-41. Heidelberg, Springer Berlin (2006).
- [4] Bu, Y., Leung, T-W., Fu, A., Keogh, E., Pei, J., and Meshkin, S., "WAT: Finding Top-K Discords in Time Series Database". In: *the 2007 SIAM International Conference on Data Mining (SDM'07)*, Minneapolis, MN, USA (2007).
- [5] Chuah, Mooi Choo, and Fen Fu, "ECG anomaly detection via time series analysis". In: *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops*, Springer Berlin Heidelberg (2007).
- [6] Lin Yi, Michael D. McCool, and Ali A. Ghorbani, "Motif and anomaly discovery of time series based on subseries join". In: *IAENG International Journal of Computer Science*, 37(3), (2010).
- [7] Buu H. T. Q. and Anh, D. T., "Time Series Discord Discovery Based on iSAX Symbolic Representation". In: *the Third International Conference on Knowledge and System Engineering (KSE 2011)*, pp. 11-18. Hanoi, Vietnam (2011).
- [8] Khanh, N. D. K. and Anh, D. T., "Time series discord discovery using WAT algorithm and iSAX representation". In: *the Third Symposium on Information and Communication Technology (SoICT'12)*, pp. 207-213. ACM New York, NY, USA (2012).
- [9] Luo, W., Gallagher, M. and Wiles, J., "Parameter-free search of time-series discord". In: *Journal of Computer Science And Technology*, vol. 28, no. 2, pp. 300-310 (2013).
- [10] Jones, M., Nikovski, D., Imamura, M. and Hirata, T., "Anomaly Detection in Real-Valued Multidimensional Time Series". In: *Proc. of 2014 ASE BIGDATA/ SOCIALCOM/ CYBERSECURITY Conference*, pp. 1-9. Stanford University (2014).
- [11] Pavel Senin, Jessica Lin, Xing Wang, Tim Oates, Sunil Gandhi, "Time series anomaly discovery with grammar-based compression". In: *18th International Conference on Extending Database Technology (EDBT)*, Brussels, Belgium (2015).
- [12] Nguyen, T. S., "Time series discord discovery based on r\*-tree". In: *Journal of Science, HCM City University of Education, Special Issue: Natural Science and Technology*, vol. 90, no. 12, pp. 133-144 (2016).
- [13] Chau, P. M., Duc, B. M., Anh, D. T., "Discord Discovery in Streaming Time Series based on an Improved HotSAX Algorithm". In: *the Ninth International Symposium on Information and Communication Technology*, pp. 24-30 (2018).
- [14] Keogh E. and Kasetty, S., "On the Need for Time Series Data Mining Benchmarks: A Survey Empirical Demonstration". In: *the 8th ACM SIGKDD Int'l Conference on Knowledge*, pp. 102-111. Edmonton, Alberta, Canada (2002).
- [15] Lin, J., Keogh, E., Wei, L., Lonardi, S., "Experiencing SAX: a novel symbolic representation of time series". In: *Journal of Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107-144 (2007).
- [16] Mueen, A., Keogh, E., Zhu, Q., Cash, S. and Westover, B., "Exact Discovery of Time Series Motifs". In: *Proc. of SIAM Int. Conf. on Data Mining*, pp. 473-484 (2009).
- [17] Keogh, E., Lonardi, S., Chiu, B., "Finding surprising patterns in a time series database in linear time and space". In: *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 550-556. New York, NY, USA (2002).

# An Adaptive Fuzzy for PMSM to Overcome the Changing Load

Nguyen Vu Quynh  
Electrical and Electronics Department  
Lac Hong University  
Dong Nai, Viet Nam  
vuquynh@lhu.edu.vn

Nguyen Hoang Huy  
Electrical and Electronics Department  
Lac Hong University  
Dong Nai, Viet Nam  
nguyenhoanhuy@lhu.edu.vn

**Abstract**—This paper proposes a robust Fuzzy PI to detect the changing of load and provide an appropriate control signal to stabilize the motor speed. First, the mathematical model of Permanent magnet synchronous motor (PMSM) and the structure of the Mamdani fuzzy controller were proved. Secondly, this controller will be combined with PI controller to adjust  $K_p$ ,  $K_i$  when the load changing. Finally, the speed response and overshoot will be discussed and compared with the PI method.

**Keywords**—PMSM, Simulation, Fuzzy PI, Adaptive Control

## I. INTRODUCTION

Permanent magnet synchronous motor (PMSM) has high efficiency, so it was widely used in industries. The motor speed controller plays a very important role. The PID controller was often used to control the speed of PMSM. However, the PID controller does not adapt when the motor load changes. The controller's quality determines the accuracy of the motor, thereby determining the quality of the whole system. With the  $K_p$  and  $K_i$  determined in advance it will increase the overshoot or response time when the motor load changes. If this problem is not studied, the results of a new control method will not be verified and cannot improve the efficiency of speed control for PMSM. The load of the motor varies, even the motor is running. The articles [1-6] also introduced a different design for the fuzzy controller. Chou [7] presented an adaptive controller based on neural networks and fuzzy processing. Although these articles' methods also achieved certain results when changing loads, but have not been tested in the case of load changes when the motor is operating. There are many types of research on intelligent controllers to control the speed of PMSM [8-9]. Moreover, quite complex neuron algorithms are not suitable for implementation on the chip.

This article proposed a method called Fuzzy PI. This method uses Mamdani fuzzy in combination with PI controller to detect the load change and give appropriate control signal to help stabilize speed. With this method, engineers will design the PMSM motor controller more effectively. It can be done on the microprocessor, increasing the applicability of actual controllers.

The rest of the paper is presented as follows: Section 2 introduces the mathematical model of PMSM and vector control method. Section 3 describes the Fuzzy PI control method. Next, section 4 presents simulated results on Simulink. Finally, some comments and assessments of the achieved results will be presented in section 5.

## II. VECTOR CONTROL

### A. The PMSM Mathematical Model

The PMSM mathematical model on the rotating reference d-q frame is used for analysis.

$$\frac{di_d}{dt} = \frac{1}{L_d} v_d - \frac{R}{L_d} i_d + \frac{L_q}{L_d} p \omega_r i_q \quad (1)$$

$$\frac{di_q}{dt} = \frac{1}{L_q} v_q - \frac{R}{L_q} i_q - \frac{L_d}{L_q} p \omega_r i_d - \frac{\lambda p \omega_r}{L_q} \quad (2)$$

In which:  $L_q$ ,  $L_d$  are the inductance on q and d axis;  $R$  is the resistance of the stator windings;  $i_q$ ,  $i_d$  are the current on q and d axis;  $V_q$ ,  $V_d$  are the voltage on q and d axis;  $\lambda$  is the permanent magnet flux linkage;  $p$  is a number of pole pairs;  $\omega_r$  is the rotational speed of the rotor.

### B. The Vector Control

The current loop control of the PMSM drive in Fig.1 is based on a vector control approach. Three phases current were being feedbacked and through vector control structure, enabling controlling current  $i_d \approx 0$ , helped to control three-phase motor similar to a DC motor. The torque of the motor is controlled via current on the q axis ( $i_q$ ).

## III. DESIGN OF FUZZY PI

### A. Block Diagram of Fuzzy PI controller

The Fuzzy PI controller's block diagram is shown in the speed loop block of Fig. 1. With  $e$  is the error between the motor's current speed and the desired speed.

$$e = \omega_r^* - \omega_r \quad (3)$$

The speed respond of the motor is a curve that closely resembles the function:

$$\omega_F(t) = \omega_r^* (1 - e^{-at}) \quad (4)$$

The equation (4) is used to compare the motor speed to give  $e_F$  error as the input of two Mamdani Fuzzy processors.



$$e_F = \omega_F - \omega_r \quad (5)$$

Fuzzy 1 helps the motor starting and quickly achieving the desired speed with different loads by accumulating errors into the  $K_p$  and  $K_i$  coefficients of the PI controller, through two IntegratorA and IntegratorB. After the speed reaches about 90% of the desired speed ( $1 - e^{-at} \geq 0.9$ ), the  $SW_1$  and  $SW_2$  will switch to the FuzzyB. During the motor operating, this unit detects the load change and helps the motor quickly stabilize with the new load by changing the  $K_p$ ,  $K_i$  coefficients of the PI to be more appropriate.

### B. Design of FuzzyA

There is one input ( $e_F$ ) and one output ( $val_1$ ). The  $val_1$  will be multiplied by the factors  $K_{pA}$  and  $K_{iA}$  and then accumulated to the  $K_p$  and  $K_i$  coefficients based on the integrators.

The fuzzyA in this article uses Mandani fuzzifier, membership function (Fig 2), product inference rule and central average defuzzifier method. In Fig. 1, the tracking error  $e_F$  and the output  $val_1$  are defined by:

$$e_F = \{\text{N2, N1, ZE, P1, P2}\}$$

$$val_1 = \{\text{DE2, DE1, ZE, IN1, IN2}\}$$

The fuzzy inference of fuzzyA:

$$\begin{aligned} \text{IF } e_F = \text{P2 THEN } val_1 &= \text{IN2} \\ \text{IF } e_F = \text{P1 THEN } val_1 &= \text{IN1} \\ \text{IF } e_F = \text{N2 THEN } val_1 &= \text{DE2} \\ \text{IF } e_F = \text{N1 THEN } val_1 &= \text{DE1} \\ \text{IF } e_F = \text{ZE THEN } val_1 &= \text{ZE} \end{aligned}$$

The output of fuzzyA:

$$val_1 = -10\mu_{N2}(e_F) - \mu_{N1}(e_F) + \mu_{P1}(e_F) + 10\mu_{P2}(e_F) \quad (6)$$

### C. Design of FuzzyB

Similar to the fuzzyA, the FuzzyB (Fig. 3) has one input ( $e_F$ ) and one output ( $val_2$ ). The  $val_2$  will be multiplied by the factors  $K_{pB}$  and  $K_{iB}$  and then accumulated to the  $K_p$  and  $K_i$  coefficients based on the integrators.

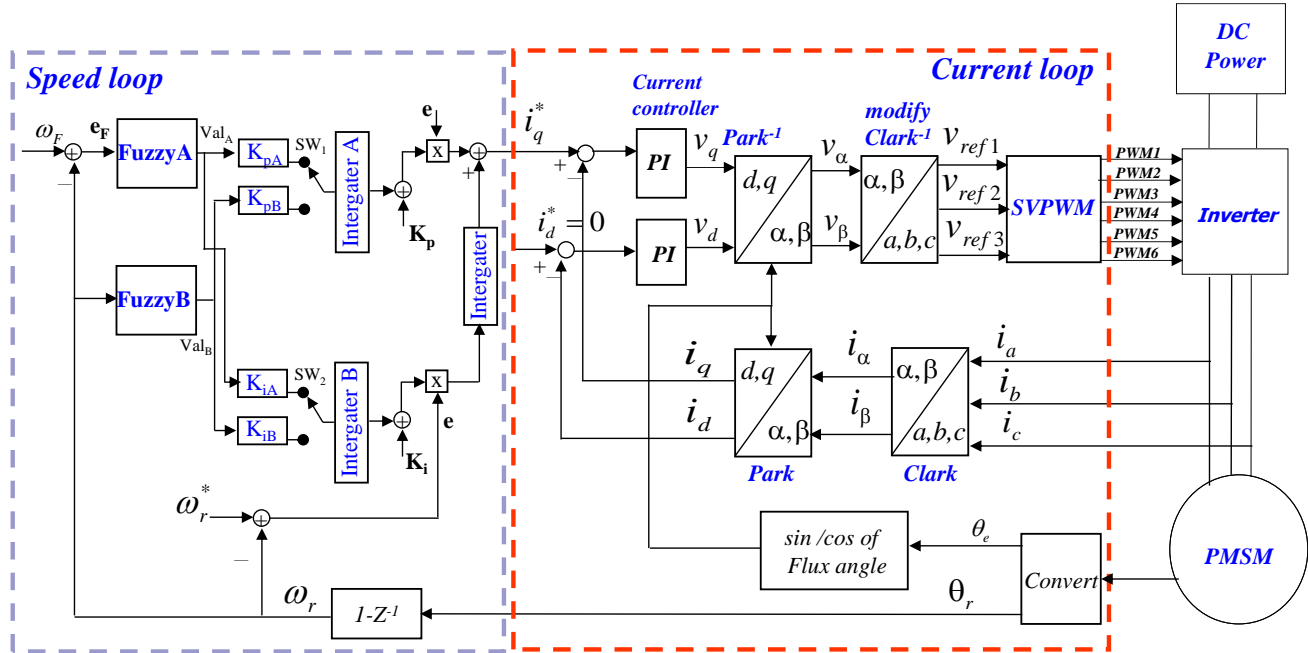


Fig. 1. The block diagram of the PMSM controller

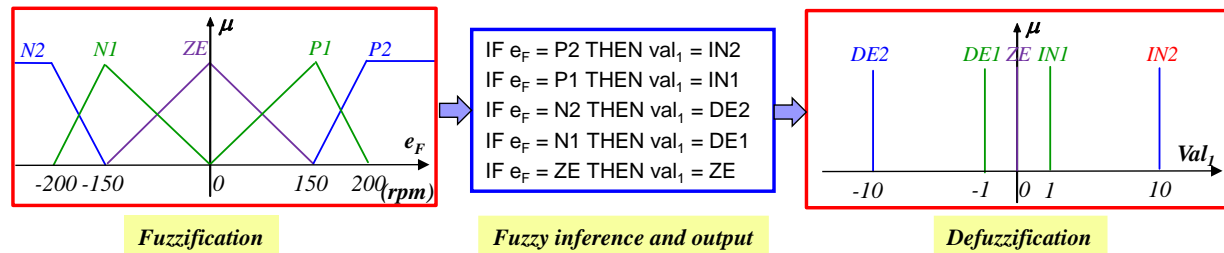


Fig. 2. The structure of the FuzzyA controller



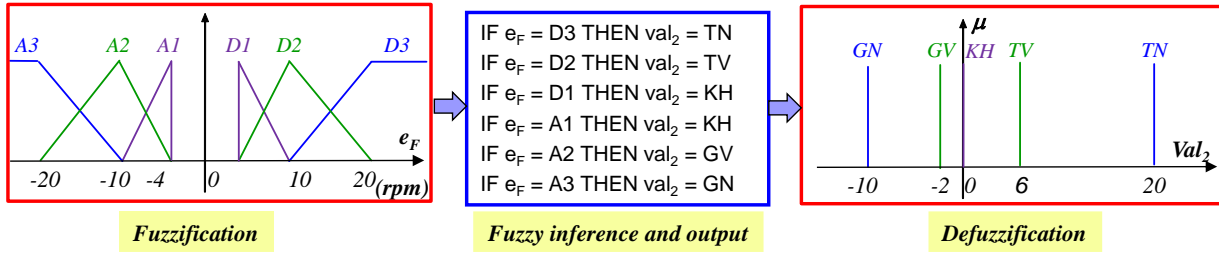


Fig. 3. The structure of the FuzzyB controller

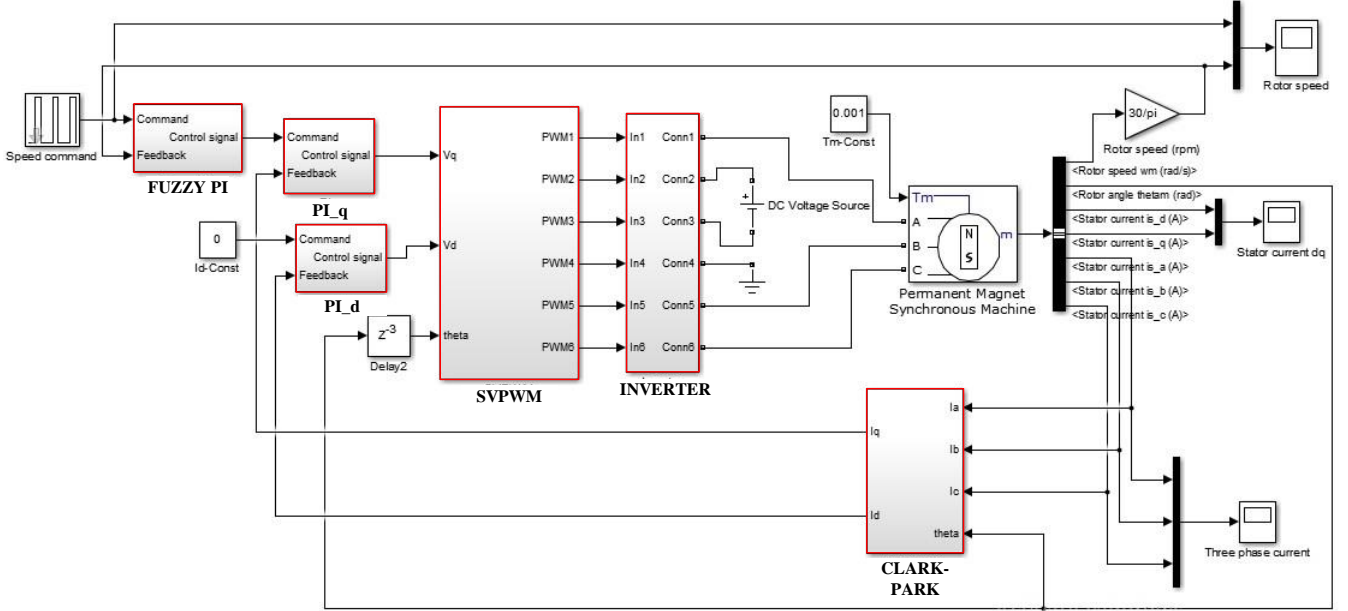


Fig. 4. The Simulink model of the PMSM speed controller

The tracking error  $e_F$  and the output of fuzzyB are defined by

$$e_F = \{A3, A2, A1, D1, D2, D3\}$$

$$val_2 = \{GN, GV, KH, TV, TN\}$$

The fuzzy inference of fuzzyB:

$$\begin{aligned} \text{IF } e_F = D3 \text{ THEN } val_2 &= TN \\ \text{IF } e_F = D2 \text{ THEN } val_2 &= TV \\ \text{IF } e_F = D1 \text{ THEN } val_2 &= KH \\ \text{IF } e_F = A1 \text{ THEN } val_2 &= KH \\ \text{IF } e_F = A2 \text{ THEN } val_2 &= GV \\ \text{IF } e_F = A3 \text{ THEN } val_2 &= GN \end{aligned}$$

The output of fuzzyB:

$$val_2 = -10\mu_{A3}(e_F) - 2\mu_{A2}(e_F) + 6\mu_{D2}(e_F) + 20\mu_{D3}(e_F) \quad (7)$$

Thus, the coefficients  $K_p$  and  $K_i$  will be accumulated according to the following formula:

If  $1 - e^{-at} < 0.9$  then:

$$K_p = K_{p0} + K_{pA} \int val_1 dt \quad (8)$$

$$K_i = K_{i0} + K_{iA} \int val_1 dt \quad (9)$$

If  $1 - e^{-at} \geq 0.9$  then:

$$K_p = K_{p0} + K_{pB} \int val_2 dt \quad (10)$$

$$K_i = K_{i0} + K_{iB} \int val_2 dt \quad (11)$$

The output of Fuzzy PI controller:

$$i_q^* = u_{out} = K_p e(t) + K_i \int e(t) dt \quad (12)$$

#### IV. SIMULATION RESULTS

The block diagram of the PMSM controller shown in Fig. 1 and Fig. 4 was its Simulink diagram. The Fuzzy PI structure is shown in Fig. 2-3. The parameters of the PMSM are given in table 1. In Fig. 4, the SimPowerSystem blockset of the Matlab/Simulink designed the PMSM and the inverter. The speed loop block in Fig.1 is designed in the Fuzzy PI block in Fig.4. The PI\_q and PI\_d (Fig. 4) are the PI controller in q and d axes, respectively. The CLARK-PARK block (Fig. 4) is the Clark, Park, Invert-Clark and Invert-Park transformation. The SVPWM block (Fig. 4) is the space vector pulse width modulation. To evaluate the effectiveness of the control method, the parameters of the motor will be changed as follows: case 1: the J and F are setup with J=0.000108, F=0.0013; Case 2: J=0.000108x2, F=0.0013x2; Case 3: J=0.000108x3, F=0.0013x3; Case 4: J=0.000108x4, F=0.0013x4.

The simulation results are shown in Fig. 5-10. They are performed with the PI and Fuzzy PI controller. The parameters

selected for PI controller are  $K_p = 2000$  and  $K_i = 1.5$ . Fuzzy PI parameters will vary according to the load with the initial value  $K_{p0} = 10$  and  $K_{i0} = 1$ . The cumulative coefficients of the Fuzzy PI controller are  $K_{pA} = 1$ ,  $K_{iA} = 0.00003$ ,  $K_{pB} = 0.6$  and  $K_{iB} = 0.000015$ . Table 2 summarizes the parameters of the controllers. Fig. 5-8 shown the result of loading in case 1, case 2, case 3, and case 4, respectively. The green line indicates the desired value, which varies from  $0 \Rightarrow 400 \Rightarrow 600 \Rightarrow 800 \Rightarrow 600 \Rightarrow 400$ . The red line is the speed response of the rotor (rpm). While the Fuzzy PI controller has not overshoot and the response time is 0.01s, the PI has overshoot up to 15%, and the response time is 0.075s when the load in case 3 and case 4.

TABLE I. PARAMETERS OF PMSM

Pole pairs	Stator phase resistance ( $\Omega$ )	Stator inductance (mH)	Inertia ( $\text{kgm}^2$ )	Friction factor (Nms)
4	1.3	6.3	0.000108	0.0013

TABLE II. PARAMETERS OF PI CONTROLLER

$K_p$	$K_i$	$K_{p0}$	$K_{i0}$
2000	1.5	10	1
$K_{p1}$	$K_{i1}$	$K_{p2}$	$K_{i2}$
1	0.00003	0.6	0.000015

Fig. 9 is simulated when the motor load in case 1 at the first 0.3s, then increases case 2 from 0.3s to 0.6s, eventually, decreases to case 1 in the remaining time. When the load increases suddenly from case 1 to case 2 the motor speed will drop by about 10 rpm for the PI while Fuzzy PI is 6 rpm and there will be no overshoot afterward. Similarly, Fig. 13 has in case 3 load from 0.3s to 0.6s. When the load increases from case 1 to case 3, the speed will decrease to 18rpm for PI and 12rpm for Fuzzy PI. The PMSM's speed response when applying the Fuzzy PI controller does not have overshoot.

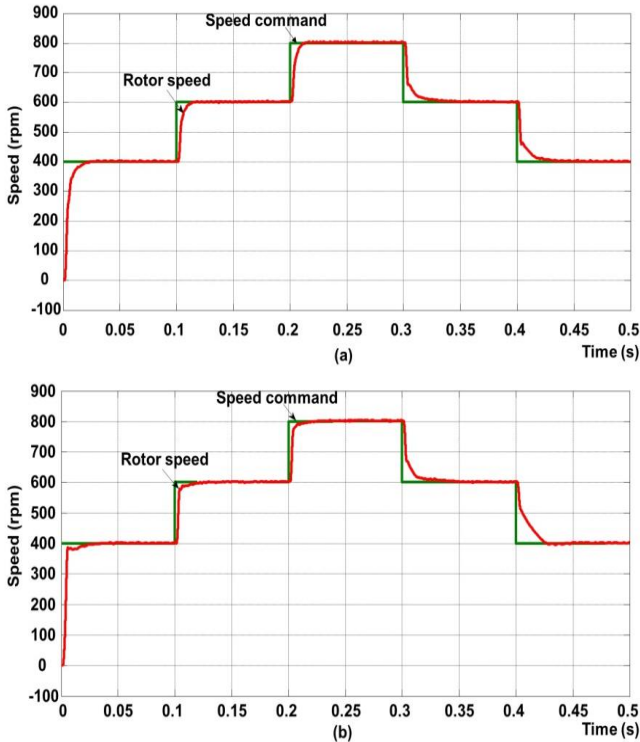


Fig. 5. The rotor speed in case 1 of the load. (a) The PI controller and (b) the Fuzzy PI controller

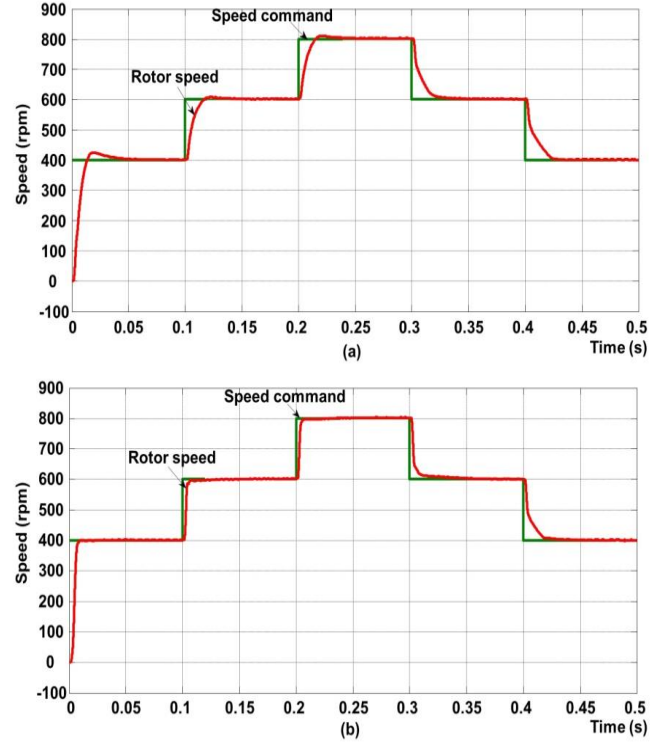


Fig. 6. The rotor speed in case 2 of the load. (a) The PI controller and (b) the Fuzzy PI controller

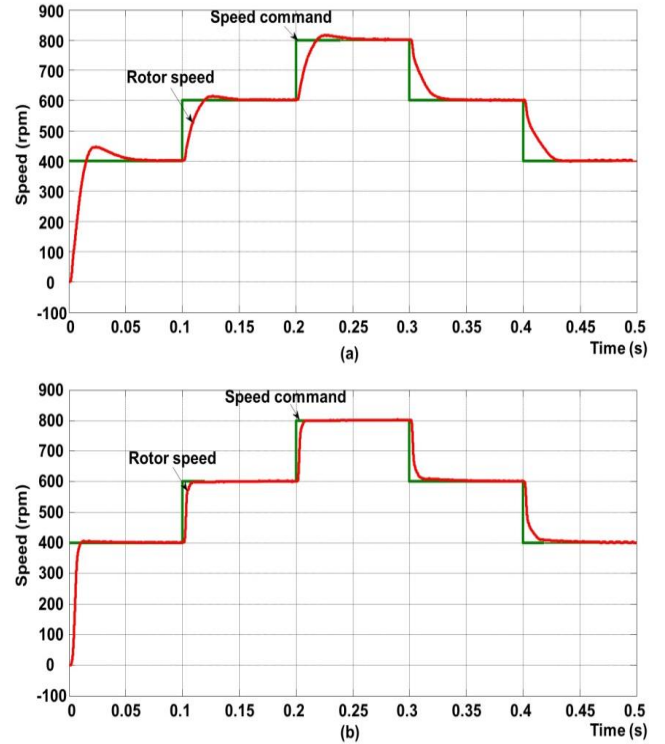


Fig. 7. The rotor speed in case 3 of the load. (a) The PI controller and (b) the Fuzzy PI controller

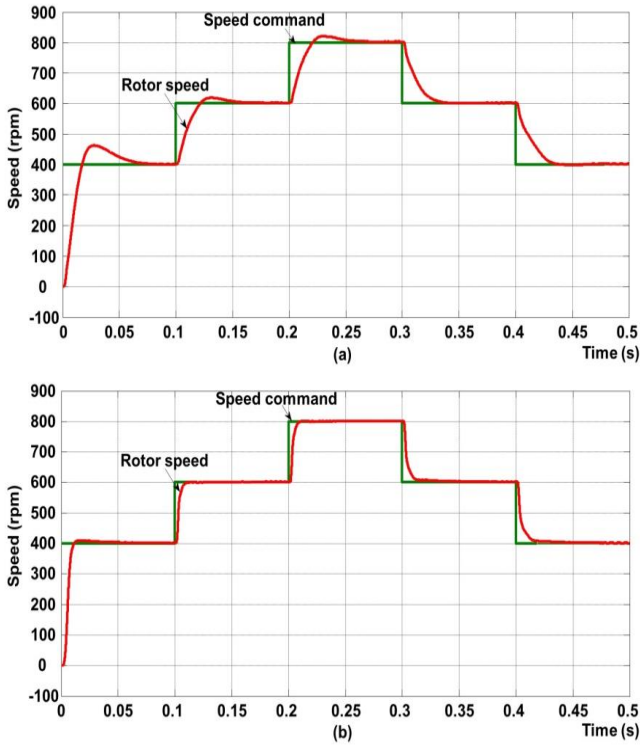


Fig. 8. The rotor speed in case 4 of the load. (a) The PI controller and (b) the Fuzzy PI controller

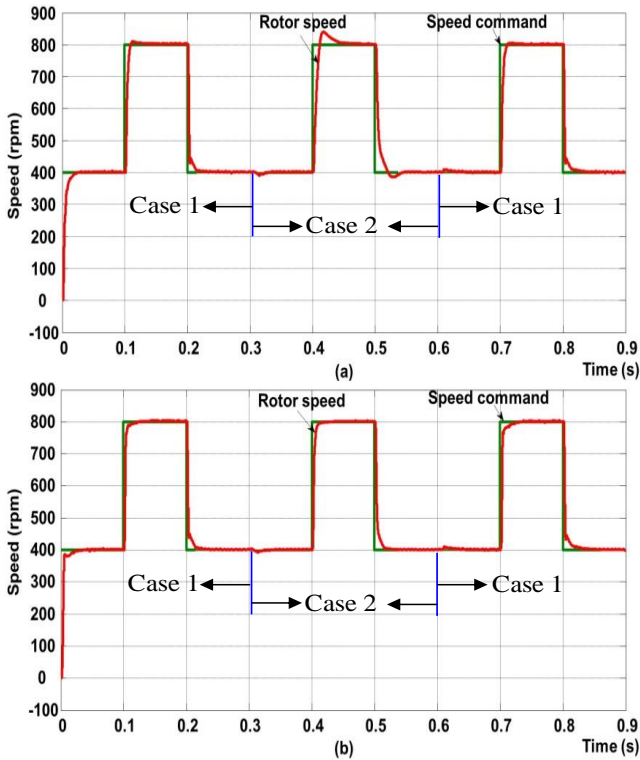


Fig. 9. The simulation results when PMSM is operated varying from case 1 to case 2. (a) The PI controller and (b) the Fuzzy PI controller

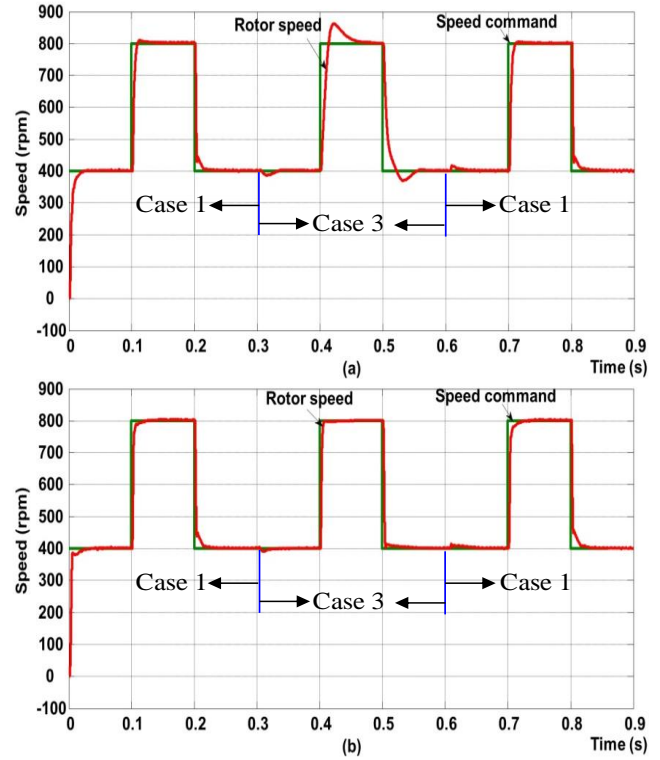


Fig. 10. The simulation results when PMSM is operated varying from case 1 to case 3. (a) The PI controller and (b) the Fuzzy PI controller

## V. CONCLUSION

Simulation results have demonstrated the ability to adapt to the changing load of the proposed method. The motor speed got a steady-state quickly, and the overshoot decrease. The fuzzy controller structure is complex, but the operations are quite simple, so it can be done on microprocessor chips. The results of the study will help the design of speed controllers for PMSM engines more flexible.

## REFERENCES

- [1] J.W. Jung, Y.S. Choi, V.Q. Leu, H.H. Choi, "Fuzzy PI-type current controllers for permanent magnet synchronous motors," *IET Electric Power Applications* 5 (1) (2011) 143–152. Doi: 10.1049/iet-epa.2010.0036
- [2] Nguyen Vu Quynh, "The Fuzzy PI Controller for PMSM's Speed to Track the Standard Model", *Mathematical Problems in Engineering*, vol. 2020, Article ID 1698213, 20 pages, 2020. Doi: 10.1155/2020/1698213
- [3] Y.S. Kung, M.H. Tsai, "FPGA-based speed control IC for PMSM drive with adaptive fuzzy control," *IEEE Transactions on Power Electronics* 22 (6) (2007) 2476–2486. Doi: 10.1109/TPEL.2007.909185
- [4] N. V. Quynh and Y. Kung, "FPGA-realization of fuzzy speed controller for PMSM drives without position sensor," 2013 International Conference on Control, Automation and Information Sciences (ICCAIS), Nha Trang, 2013, pp. 278–282. doi: 10.1109/ICCAIS.2013.6720568
- [5] Chih-Min Lin, Tien-Loc Le, Tuan-Tu Huynh, "Self-evolving function-link interval type-2 fuzzy neural network for nonlinear system identification and control," *Neurocomputing*, Volume 275, 2018, Pages 2239–2250. Doi: 10.1016/j.neucom.2017.11.009
- [6] Atif Iqbal, Haitham Abu-Rub & Hazem Nounou, "Adaptive fuzzy logic-controlled surface mount permanent magnet synchronous motor drive," *Systems Science & Control Engineering*, 2014, Pages 465–475. doi: 10.1080/21642583.2014.915203
- [7] Hsin-Hung Chou, Ying-Shieh Kung, Nguyen Vu Quynh, Stone Cheng, "Optimized FPGA design, verification and implementation of a neuro-fuzzy controller for PMSM drives," *Mathematics and Computers in Simulation*, Volume 90, April 2013, Pages 28–44, ISSN 0378-4754, Doi: 10.1016/j.matcom.2012.07.012

- [8] Zhang Chunmei, Liu Heping, Chen Shujin, Wang Fangjun, "Application of neural networks for permanent magnet synchronous motor direct torque control," *Journal of Systems Engineering and Electronics*, Volume 19, Issue 3, 2008, Pages 555-561. doi: 10.1016/S1004-4132(08)60120-6.
- [9] S. Li, H. Won, X. Fu, M. Fairbank, D. C. Wunsch and E. Alonso, "Neural-Network Vector Controller for Permanent-Magnet Synchronous Motor Drives: Simulated and Hardware-Validated Results," in *IEEE Transactions on Cybernetics*. doi: 10.1109/TCYB.2019.2897653

# Crack Patterns in Direct Tension and Flexure of Ultra-High-Performance Fiber-Reinforced Concrete with Scale Effect

Duy-Liem Nguyen  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
liemnd@hcmute.edu.vn

Vu-Tu Tran  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tutv@hcmute.edu.vn

Huynh-Tan-Tai Nguyen  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tainhtt@hcmute.edu.vn

Min-Kyoung Kim  
Department of Civil and  
Environmental Engineering,  
Sejong University  
Seoul, Republic of Korea  
mkkim9112@naver.com

**Abstract**— The crack patterns of ultra-high-performance fiber-reinforced concrete (UHPFRC) with size effect in tension and flexure were experimentally investigated. The investigated UHPFRC was blended 1.0 vol.% long twisted with 1.0 vol.% short smooth steel fibers. Flexural specimens and tensile specimens were designed with different scales of sizes as follows: 25×50×125, and 50×100×250 mm (thickness×width×gauge length) for the tensile specimens, 50×50×150, 100×100×300, 150×150×450 mm (width×depth×span length) for the flexural specimens. All tested specimens exhibited multiple micro-cracks during work-hardening responses. There was a clear scale effect on number of cracks appearing in UHPFRC specimens under tension and flexure, in addition to scale effect on strength, deformation, toughness. Besides, the model and the degree of scale effect on micro-cracks were shown and discussed.

**Keywords**— Scale effect, Crack pattern, Strain-hardening, Deflection-hardening, Micro-cracks

## I. INTRODUCTION

Ultra-high-performance concrete (UHPC) has been categorized as one of advanced cementitious materials with ultra-high strength more than 150 MPa and superior durability due to its densified microstructure [1,2]. However, UHPC still reveals its brittle nature, similar to normal concrete [3,4]. The adding fibers in UHPC will form a mixture called ultra-high-performance fiber-reinforced concretes (UHPFRC). In both direct tension and bending, UHPFRC could demonstrate its high ductility with multiple micro-cracks during work hardening [5,6], although UHPFRC still revealed the brittle failure in compression [7].

Clearly, the work hardening mechanism is the most important property of UHPFRC producing its high tensile strength, high strain capacity as well as toughness, and, the observed crack patterns with multiple micro-cracks were the main proofs. However, the scale effect on micro-cracks of UHPFRC, in forgoing study works [3,4], was not focused yet. This was really a gap information on mechanical properties of UHPFRC. Understanding the cracking behaviors of UHPFRC

can help widen practical applications of UHPFRC in structural members such as long bridges, tall buildings, etc.

The above situation was the motivation for the investigation presented in this study work. The main objective is to investigate the scale effect on micro-cracks of UHPFRCs in tension and flexure.

## II. EXPERIMENT

### A. Materials and preparation of specimens

Fig. 1 shows the photos of long twisted and short smooth steel fibers used in this work. The investigated UHPFRC was added a blend of 1.0 vol.% long twisted and 1.0 vol.% short smooth steel fibers. Tables I&II provide the composition of UHPC and fiber properties, respectively. Flexural specimens and tensile specimens were designed with different scales of sizes as follows: 25×50×125 mm and 50×100×250 mm (thickness×width×gauge length) for the tensile specimens, 50×50×150 mm, 100×100×300 mm, 150×150×450 mm (width×depth×span length) for the flexural specimens. The tensile specimens were bell-shaped and reinforced with steel wire meshes at two ends of specimens to prevent failure outside the gauge length.

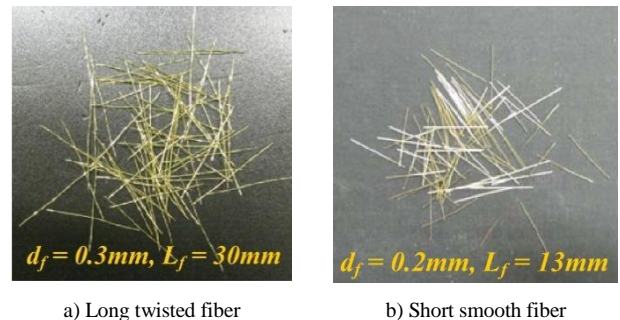


Fig. 1. Photos of the hybrid fibers used



TABLE I. COMPOSITION OF UHPC

Cement (Type I)	Silica fume	Silica sand	Silica powder	Superplasticizer	Water
1.00	0.25	1.10	0.30	0.067	0.20

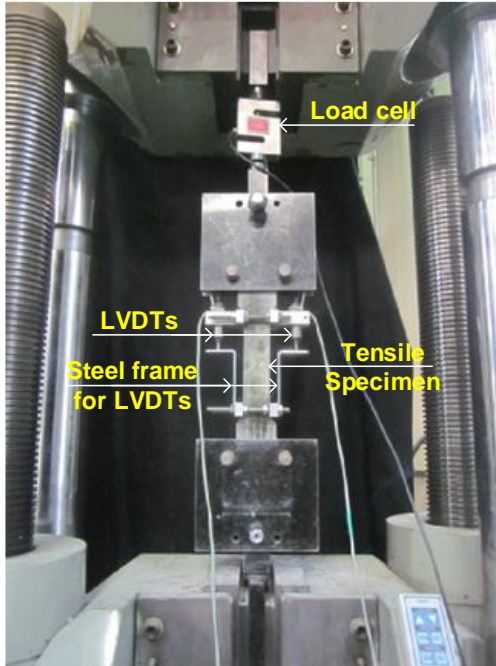
TABLE II. FIBER PROPERTIES

Fiber type	Diameter (mm)	Length (mm)	Aspect ratio (L/D)	Tensile strength (MPa)
Long twisted	0.3	30	100	2428
Short smooth	0.2	13	65	2788

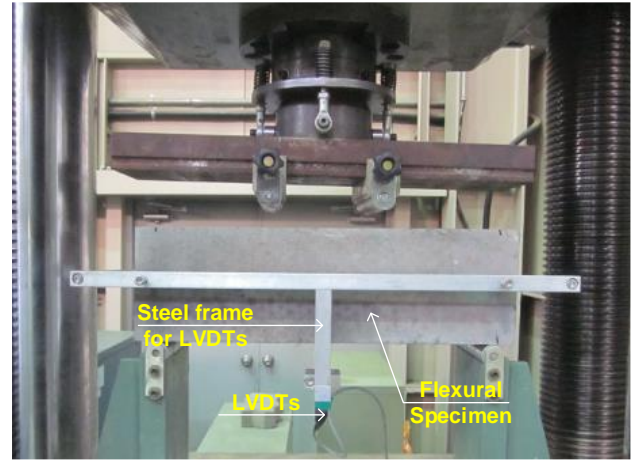
The mixing detail of UHPFRC mixture could be referred to previous studies [3,4]. All the specimens were cured in hot water,  $90 \pm 2$  °C for three days and kept more one day in room temperature water. Next, they were removed out of the water and dried at room temperature. Later, they were sprayed on specimens' surfaces with two or three layers of polyurethane for crack exposure. The testing age of 14 days was applied for both tensile and flexural specimens.

### B. Experiment setup

All specimens were tested by using a universal test machine with applied displacement speed of 1 mm/min. The frequency of data acquisition under compression tests was 1 Hz. Fig. 2 presents the experiment setup for direct tension and flexure. The linear variable differential transformers (LVDTs) were attached on tensile and flexural specimens to measure displacement or deflection, respectively. The flexural specimens were examined under the four-point bending test (4PBT). Each series was tested by using at least three specimens and used average value for evaluating.



a) Direct tensile test



b) Bending test

Fig. 2. Experiment setup

The direct tensile stress ( $\sigma$ ) and bending stress ( $f$ ) would be computed using Eq. (1) and Eq. (2), respectively.

$$\sigma = P / A \quad (1)$$

$$f = PS / (bh^2) \quad (2)$$

Where  $P$  is the applied load,  $A$  is the sectional area of the tensile specimen;  $S$  is span length,  $b$  and  $h$  are the width and depth of the flexural specimen.

## III. TEST RESULT AND DISCUSSION

### A. Tensile and flexural behaviors of UHPFRCs

Fig. 3 shows the tensile stress versus strain response curves while Fig. 4 displays the bending stress versus normalized deflection response curves of the UHPFRC. As shown in Fig. 3, both tensile series revealed the strain-hardening behaviors although the big tensile specimen exhibited the low strain capacity. Similarly, three flexural series in Fig. 4 exhibited deflection-hardening behaviors regardless of specimen size.

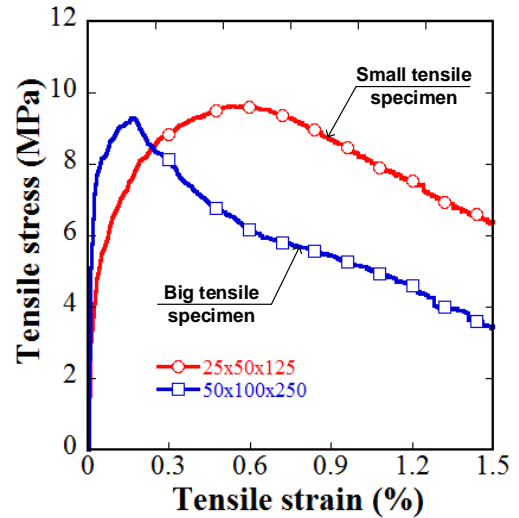


Fig. 3. Tensile behaviors of the UHPFRC

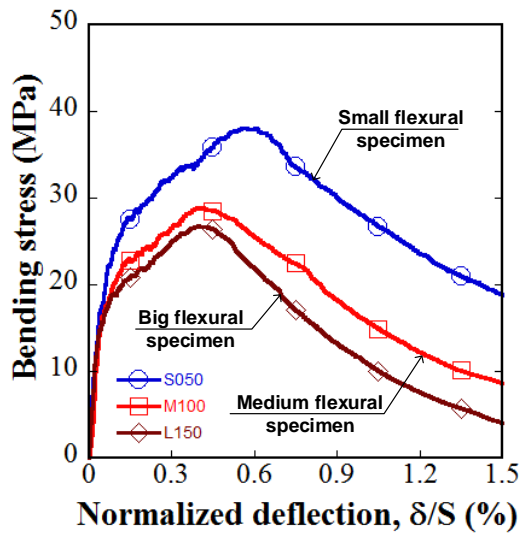


Fig. 4. Flexural behaviors of the UHPFRC

TABLE III. TENSILE PARAMETERS

Series	Post-cracking Tensile Strength (MPa)	Tensile Strain Capacity (%)	Tensile Toughness (MPa.%)
Small	9.77	0.564	4.68
Big	8.92	0.174	1.52

TABLE IV. FLEXURAL PARAMETERS

Series	Bending Strength (MPa)	Normalized Deflection Capacity (%)	Bending Toughness (MPa.%)
Small	38.91	0.595	18.00
Medium	29.10	0.424	9.69
Big	26.99	0.414	8.89

The scale effects on tensile parameters of the UHPFRC (including post-cracking tensile strength, tensile strain capacity, tensile toughness) and flexural parameters of the UHPFRC (including bending strength, normalized deflection capacity, bending toughness) were explored and presented in Tables III & IV, respectively. Here, the toughness was defined as the area under the stress versus strain (or normalized deflection) curve up to the peak. Under tension, there was a little scale effect on tensile strength (strength ratio of the small to the big  $\sim 1.1$ ) whereas a significant scale effect on tensile strain capacity was found (strain capacity ratio of the small to the big  $\sim 3.24$ ). Under flexure, there were important scale effects on bending strength and normalized deflection capacity (parameter ratio of the small to the big  $\sim 1.44$ ). The most significant scale effect was found for toughness parameter, in both tension and flexure, because the scale effect on toughness was totally resulted from scale effect on strength and scale effect on deformation.

#### B. Crack patterns in tension and flexure of UHPFRC

Figs. 5 & 6 show the crack patterns of the UHPFRC under tension and flexure, respectively. The law of crack pattern was as follows: crack spacing increased (or crack number within a unit length decreased) when the specimen size increased

regardless of loading type, i.e., the scale effect on crack spacing was observed. In addition, the crack spacing in tension seemed larger than that in flexure. It was noticed that the lower crack spacing revealed the better cracking resistance of the materials. Table V provides the counted number of crack of all tested specimens and their corresponding crack spacings. Fig. 7 displays the comparative crack number within 100 mm under tension and flexure of UHPFRC. The scale effect on crack number was significant under tension (crack number ratio of the small to the big  $\sim 2.07$ ) but less significant under flexure (crack number ratio of the small to the big  $\sim 1.24$ ).



a) Small tensile specimen, crack spacing=6.41 mm



b) Big tensile specimen, crack spacing=13.30 mm

Fig. 5. Crack patterns under direct tension of UHPFRC



a) Small flexural specimen, crack spacing=3.85 mm



b) Medium flexural specimen, crack spacing=4.55 mm



c) Big flexural specimen, crack spacing=4.69 mm

Fig. 6. Crack patterns under flexure of UHPFRC



TABLE V. CRACK NUMBER COUNTED WITHIN 100 MM UNDER TENSION AND FLEXURE

Notation of series	Spe.	Number of cracks	Crack number within 100 mm	Average crack number	Standard deviation
Tension-small (25×50×125) Gage length= 125	SP1	19	15.20	15.6	1.94
	SP2	21	16.80		
	SP3	23	18.40		
	SP4	16	12.80		
	SP5	20	16.00		
	SP6	18	14.40		
Tension-big (50×100×250) Gage length= 250	SP1	17	6.80	7.52	0.59
	SP2	18	7.20		
	SP3	21	8.40		
	SP4	19	7.60		
	SP5	19	7.60		
Flexure-small (50×50×150) Gage length= 50	SP1	12	24.00	26.67	2.31
	SP2	14	28.00		
	SP3	14	28.00		
Flexure-medium (100×100×300) Gage length= 100	SP1	22	22.00	22.00	1.00
	SP2	21	21.00		
	SP3	23	23.00		
Flexure-big (150×150×450) Gage length= 150	SP1	32	21.33	21.56	1.68
	SP2	35	23.33		
	SP3	30	20.00		

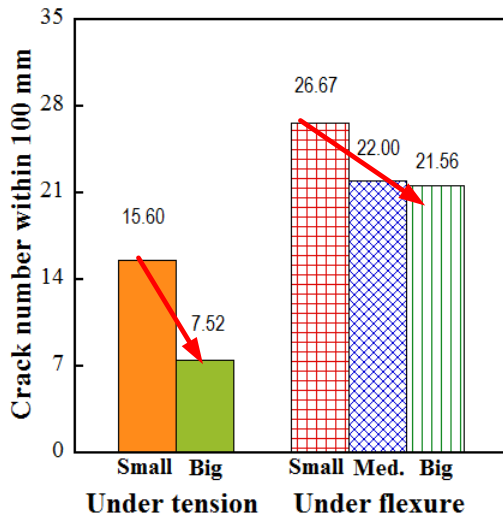
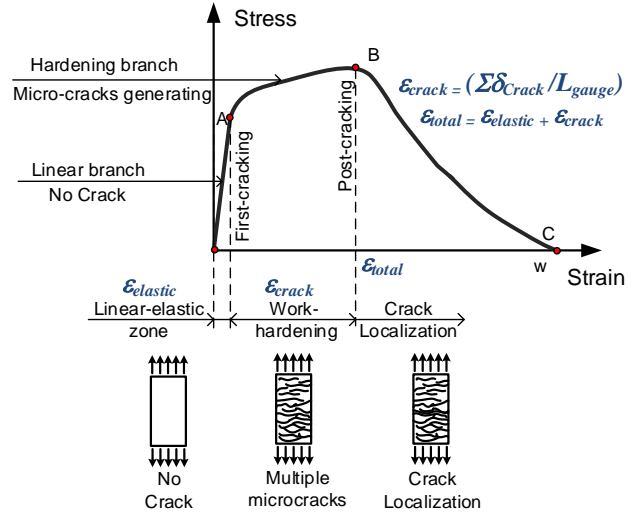


Fig. 7. Crack number within 100 mm under tension &amp; flexure of UHPFRC

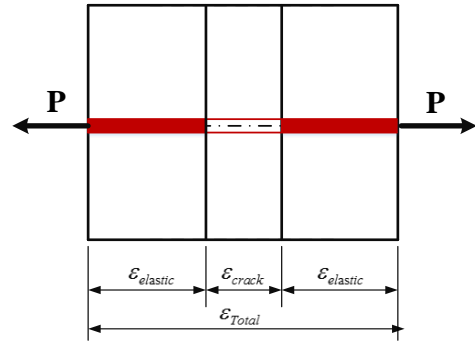
### C. Model describing crack patterns of UHPFRC under tension & flexure

Fig. 8a shows a typical stress-strain response of UHPFRC under direct tension and flexure, it is noticed that the bending stress and bending strain would be analyzed at beam bottom. As shown in Fig. 8a, the total of strain ( $\epsilon_{total}$ ) in the gauge length ( $L_{gauge}$ ) would include the elastic strain ( $\epsilon_{elastic}$ ) and crack strain ( $\epsilon_{crack}$  = total of crack widths/ gauge length). Figs. 8b and 8c display the models for micro-crack patterns under tension and flexure, respectively. In Figs. 8b & 8c, the models were assumed one crack with its width being sum of all micro-crack widths. The models for micro-crack patterns of

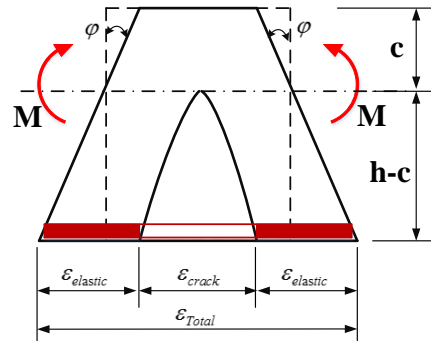
UHPFRC revealed that the scale effect on crack patterns was associated with scale effect on strain (or normalized deflection) rather than scale effect on stress.



a) Typical stress – strain response of a work-hardening UHPFRC



b) Under direct tension



c) Under flexure

Fig. 8. Models for micro-crack patterns of UHPFRC

### D. Weibull modulus for crack spacing of UHPFRC

To explain the size effect on the strength of material, two major approaches have been developed: the statistical and deterministic approaches. While Weibull's theory is a representation in statistical approach: a larger specimen has a weaker strength because it has more elements in specimens resulting higher chance of failure. A Weibull weakest-link model is illustrated as a chain in tension and as a system of a brick-chain combination in flexure [8], as shown in Fig. 9. The

chain or the system of brick-chain will break as the weakest element is failed, and, the more elements in the system, the higher probability containing the defect element.

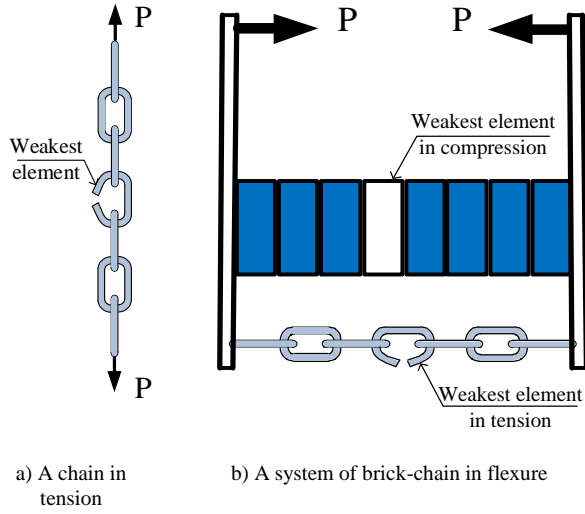


Fig. 9. Weibull weakest-link models in tension and flexure

Although the Weibull's theory was developed for strength, it could be applied for other mechanical parameters related to size effect, including parameter of crack number in a unit specimen length.

Considering the crack number ( $CN$ ) of two different sizes of specimens,  $CN_1$  and  $CN_2$ , with effective volumes  $V_{E1}$  and  $V_{E2}$ , respectively, their correlation is given by Eq. (3) according to Weibull's theory. In Eq. (3),  $V_E$  is the effective volume of the specimen according to loading condition, and  $m$  is the Weibull modulus that is considered to be a material parameter describing the size effect as well as the brittleness of material.

$$\frac{CN_1}{CN_2} = \left\{ \frac{V_{E2}}{V_{E1}} \right\}^{\frac{1}{m}} \quad (3)$$

As described in Eq. (3), if  $V_{E1}$  is greater than  $V_{E2}$ , the crack number  $CN_1$  of the larger specimen is greater than the crack number  $CN_2$  of the smaller specimen. And, material with the lower value of  $m$  is more brittle. The probability of failure,  $P_f(CN)$ , is proposed using Eq. (4) and then written into Eq. (5). In these equations,  $CN$  is the crack number and  $CN_0$  is the scale parameter.

$$P_f(CN) = 1 - \exp \left[ -V_E \left( \frac{CN}{CN_0} \right)^m \right] \quad (4)$$

$$\ln \left\{ \ln \left[ \frac{1}{(1 - P_f(CN))} \right] \right\} = m \ln(CN) + \ln(V_E) - m \ln(CN_0) \quad (5)$$

$$y_i = mx_i + b \quad (6)$$

$P_f(CN)$  is  $i/(n+1)$ ,  $n$  is the number of analyzed specimens while  $i$  is the order of crack number:  $CN_1 \leq CN_2 \leq \dots \leq CN_i \leq \dots \leq CN_n$ . Using the least square method on

the linearized regression model given by Eq. (6), the value of the Weibull modulus,  $m$ , can be estimated using Eq. (7) with the data:  $y_i = \ln \left\{ \ln \left[ \frac{1}{(1 - i/(n+1))} \right] \right\}$  and the corresponding

$$x_i = \ln(CN_i) \quad , \quad \text{the average value} \quad y_{av} = \frac{1}{n} \sum_i y_i \quad \text{and}$$

$$x_{av} = \frac{1}{n} \sum_i x_i \quad .$$

$$m = \frac{\sum_i [(y_i - y_{av})(x_i - x_{av})]}{\sum_i (x_i - x_{av})^2} \quad (7)$$

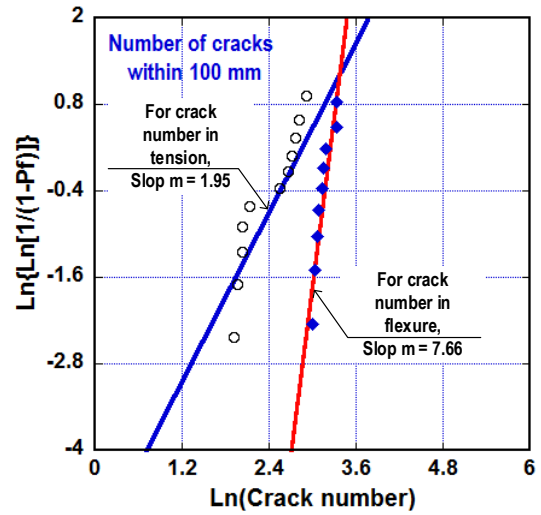


Fig. 10. Using the least-squares method to obtain the Weibull modulus of crack number

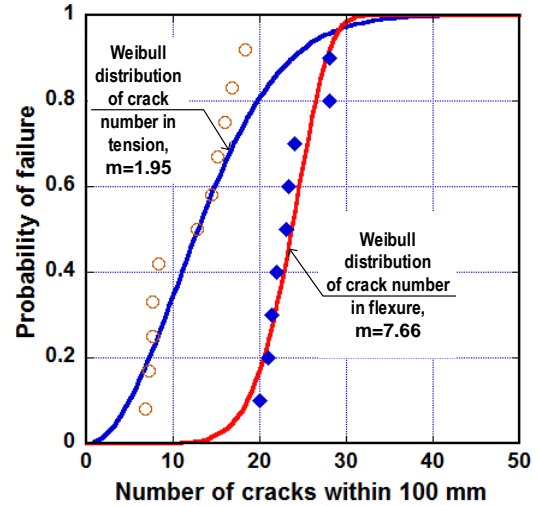


Fig. 11. Weibull distribution of crack number

Fig. 10 presents the least-squares method to obtain the Weibull modulus of crack number in tension,  $m_{CN-ten}$ , and crack number in flexure,  $m_{CN-ben}$ . The  $m_{CN-ten}$  of UHPFRC was analyzed using 11 tested specimens with two different scales while the  $m_{CN-ben}$  of UHPFRC was analyzed using 9 tested specimens with three different scales. Fig. 11 shows the Weibull distribution of crack numbers under tension and

flexure of UHPFRC. The Weibull modulus were obtained as follows:  $m_{CN-ten} = 1.95 < m_{CN-ben} = 7.66$ . This meant the scale effect on crack number in tension was more significant than that in flexure. Compared to Weibull modulus for strength as follows:  $m = 8.5$  for tensile strength [3] and  $m = 9.6$  for flexural strength [4], the Weibull modulus for crack number was rather smaller, i.e., the scale effect on crack number was more significant than scale effect on strength, in both tension and flexure of UHPFRC.

#### IV. CONCLUSIONS

Some conclusions could be derived from this investigation:

- The investigated UHPFRCs produced multiple micro-cracks during work-hardening in both tension and flexure. And, the scale effect on crack number was observed, in addition to scale effect on strength, deformation (or normalized deflection) and toughness.
- The presented models for micro-crack patterns of UHPFRC revealed that the scale effect on crack patterns was associated with scale effect on strain (or normalized deflection) rather than scale effect on stress.
- The Weibull modulus of UHPFRC for crack number within a unit gauge length under tension was smaller than that under flexure, i.e., the clearer scale effect on crack was observed in tension.

#### ACKNOWLEDGMENT

This research was supported by Ho Chi Minh City University of Technology and Education, the authors are grateful to the sponsor. The opinions expressed in this paper

are those of the authors and do not necessarily reflect the views of the sponsor.

#### REFERENCES

- [1] P. Rossi, A. Antonio, E. Parant, P. Fakhri, "Bending and compressive behaviors of a new cement composite," *Cement Concrete Res.*, 35(1), 2005, pp. 27–33.
- [2] Graybeal, B. and Davis, M., "Cylinder or cube: strength testing of 80 to 200 MPa (11.6 to 29 ksi) Ultra-High-Performance-Fiber-Reinforced Concrete," *ACI Mater. J.*, 105(6), 2008, pp. 603–9.
- [3] D. L. Nguyen, G. S. Ryu, K. T. Koh, D. J. Kim, "Size and geometry dependent tensile behavior of ultra-high-performance fiber-reinforced concrete," *Composites: Part B*, 58 (2014): pp. 279-292.
- [4] D. L. Nguyen, D. J. Kim, G. S. Ryu, K. T. Koh, "Size Effect on flexural behavior of ultra-high-performance hybrid fiber-reinforced concrete," *Composites: Part B*, 45 (2013), pp. 1104-1116.
- [5] K. Wille, D. J. Kim, A. E. Naaman, "Strain hardening UHP-FRC with low fiber contents," *Mater Struct*, 44 (2011):583-98.
- [6] D. J. Kim, S. H. Park, G. S. Ryu, K. T. Koh, "Comparative flexural behavior of hybrid ultra high performance fiber reinforced concrete with different macro fibers," *Constr. Build. Mater.* 25 (2011) 4144–4155.
- [7] D. L. Nguyen and D. J. Kim, "Sensitivity of various steel-fiber types to compressive behavior of ultra-high-performance fiber-reinforced concretes," in *Proceedings of AFGC-ACI-fib-RILEM International Symposium on Ultra-High Performance Fibre-Reinforced Concrete, UHPFRC 2017, PRO 106 - RILEM Publications*, October 2-4, 2017, Montpellier, France, pp 45-52.
- [8] D. L. Nguyen, D. K. Thai, T. T. Ngo, T. K. Tran, T. T. Nguyen, "Weibull modulus from size effect of high-performance fiber-reinforced concrete under compression and flexure," *Constr. Build. Mater.* 2019, 226, 743–758.

# Indirect Tensile Strengths of Crushed-Sand Concretes in Correlation with Its Compressive Strength

Duy-Liem Nguyen  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
liemnd@hcmute.edu.vn

Vu-Tu Tran  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tutv@hcmute.edu.vn

Thi-Ngoc-Han Vuong  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
hanvtn@hcmute.edu.vn

Tien-Tho Do  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thodt@hcmute.edu.vn

**Abstract**— The correlation between indirect tensile strengths and compressive strengths of the crushed-sand concretes was experimentally investigated. The crushed-sand concrete called in this paper was termed as the concrete using crushed-sand to replace traditional river sand in producing concrete. Six types of the crushed-sand concretes were examined with different compositions. The indirect tensile test included three-point bending test and cylinder splitting test as follows: dimensions of bending specimen was 100×100×300 mm (width×depth×span-length) while that of splitting specimen was 100×200 mm (diameter×length). The compressive test was used cube specimen of 150×150×150 mm. The testing results were obtained then comparatively evaluated.

**Keywords**— *Crushed Sand, Fine Aggregate, Fly Ash, Environmental Impact, Landslide*

## I. INTRODUCTION

Concrete together with steel are the most dominant materials of construction industry. In Vietnam, the traditional concrete has been used natural river sand as fine aggregate in its composition. Recently, there has been a great demand in the consumption of concrete for a lot of construction activities, this leads to a large amount of sand used. However, the excessive exploit of the river sand has been caused environment problem: landslides occurred more commonly and seriously, especially in Mekong Delta river zone, Southern Vietnam, as illustrated in Fig. 1 [1,2]. Therefore, finding a substitute to natural river sand has become urgent and necessary. Crushed-sand is an artificial material made from waste production in coarse aggregate manufacturing process. In the previous study [3], the crushed sand manufactured in An Giang, a province of Mekong Delta river zone, was reported to be practicably replaced the natural river sand: it could satisfy technical requirements for producing concrete with lower cost. Furthermore, crushed-sand concrete was also used fly ash, another industrial wastes from coal-fired power stations, as partial cement replacement material. Consuming fly ash would contribute to recycling this industrial waste. The utilization of crushed sand and fly ash is highly expected to

bring much benefits in reducing the construction cost and minimizing negative environmental impact.

The aim of this research is to clear the mechanical properties of the crushed-sand concretes, and, the research objective is to correlate the indirect tensile strengths and compressive strengths of them. A better understanding of the mechanical properties of the crushed-sand concrete will help civil engineers to apply this concrete type in structural members with proper design and high reliability.



Figure 1. Excessive exploit of natural river sand causing landslide problem.

## II. DIRECT AND INDIRECT TENSILE STRENGTH OF CONCRETE

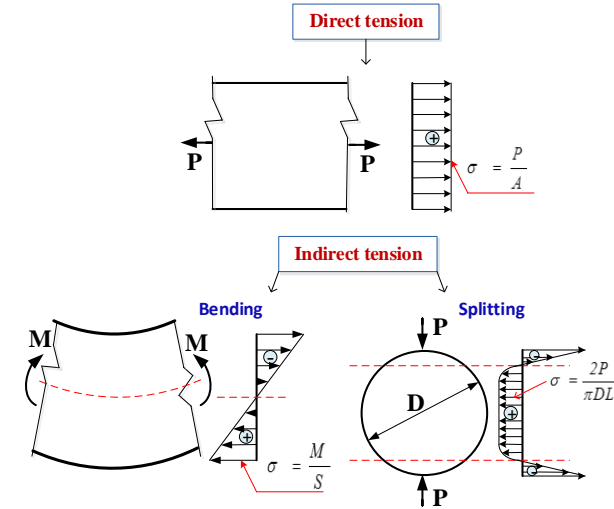


Figure 2. Distributions of stress on specimen section using various tensile tests.

In practice, normal concrete is unusually tested under direct tensile loading owing to hard measuring data of loading and specimen elongation, the main cause is from pretty poor tensile strength of concrete. To overcome this problem, bending test and splitting test, considered as indirect tensile tests, have been recommended for testing concrete. The different distributions of stress on cross-section was displayed in Fig. 2 for direct and indirect tensile tests. As described in Fig. 2, both compressive stress and tensile stress were existed under indirect tension, however, under direct tension, only tensile stress appeared [4,5]. The presence of the compressive phase in indirect tension is really the key factor producing the higher values of failure strengths. In ACI-318 [6], ratio between tensile strength and square root of compressive strength of concrete was a constant, given by Eqs. 1-3. The direct and indirect tensile strengths were obtained using Eqs. 4-6. The order of the scale coefficients defined in Eqs. 1-3 of normal concrete are adopted by ACI-318 as follows:  $K_o = 0.33 < K_{SPL} = 0.52 < K_{MOR} = 0.63$ .

For crushed-sand concrete, the scale coefficients,  $K_{SPL}$  and  $K_{MOR}$ , would be experimentally detected.

$$\sigma_o = K_o \sqrt{f'_c} \quad (1)$$

$$f_{MOR} = K_{MOR} \sqrt{f'_c} \quad (2)$$

$$f_{SPL} = K_{SPL} \sqrt{f'_c} \quad (3)$$

$$\sigma_o = P_{\max} / A \quad (4)$$

$$f_{MOR} = M_{\max} / S = 1.5 P_{\max} L_{span} / (bh^2) \quad (5)$$

$$f_{SPL} = 2P_{\max} / (\pi DL) \quad (6)$$

where  $f'_c$  is the compressive strength using cylinder specimen of 150x300 mm;  $\sigma_o$ ,  $f_{MOR}$  and  $f_{SPL}$  are the direct tensile strength, three-point bending strength and

splitting strength of traditional concrete, respectively.  $P_{\max}$  is the maximum measured load;  $A$  is the cross-section area of the direct tensile specimen.  $h$  and  $b$  are height and width of cross-section, respectively, of the bending specimen.  $L_{span}$  is the span length,  $D$  and  $L$  are the diameter and length of the splitting specimen, respectively.

## A. Materials and specimen preparation

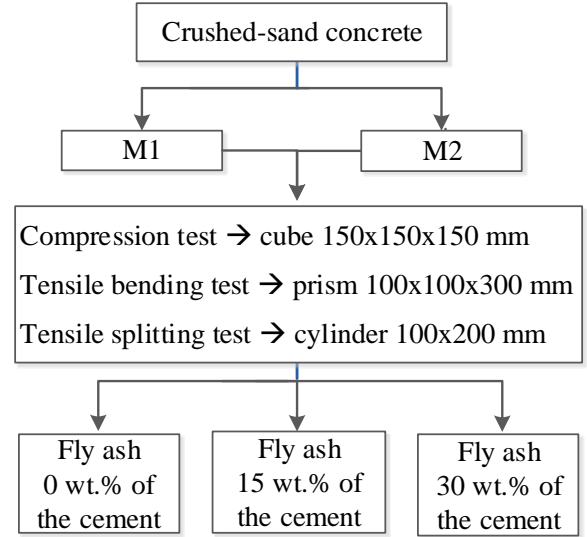


Figure 3. Test flowchart.

The test flowchart was presented in Fig. 3 while the compositions and compressive strength of the crushed-sand concretes were provided in Table 1. As shown in Fig. 3 and Table 1, there were two basic matrixes named M1 and M2. For each matrix, the fly ash amounts were changed as follows: 0%, 15% and 30 wt.% of the cement, total of cement and fly ash weight was fixed for each matrix. The particle size fractions of crushed sand was shown in Table 2. Both crushed sand and aggregate were produced by An Giang Stone Exploitation & Processing One Member Company [7]. The compressive specimens were cube-shaped with dimensions of 150x150x150 mm, the bending specimens were prism-shaped with dimensions of 100x100x300 mm and splitting specimens were cylinder-shaped with dimensions of 100x200 mm. All tested specimens were cured for 24 days in water with temperature 28-32°C average. The specimens were experimented at 28-day age under dry condition. The used cement was PCB40 of An Giang ACIFA.

TABLE I. COMPOSITIONS OF THE CONCRETES

Name of series	Cement (kG)	Fly ash (kG)	Aggregate size 1x2 (kG)	Crushed sand (kG)	Water (kG)	Comp. strength (MPa)
M1-T00	320 (1.00)	0 (0)	1150 (3.59)	715 (2.23)	215 (0.67)	27.19
M1-T15	272 (0.85)	48 (0.15)	1150 (3.59)	715 (2.23)	215 (0.67)	36.87
M1-T30	224 (0.70)	96 (0.30)	1150 (3.59)	715 (2.23)	215 (0.67)	18.17
M2-T00	445 (1.00)	0 (0)	1070 (2.40)	651 (1.46)	234 (0.53)	34.48
M2-T15	378 (0.85)	67 (0.15)	1070 (2.40)	651 (1.46)	234 (0.53)	43.13
M2-T30	312 (0.70)	133 (0.30)	1070 (2.40)	651 (1.46)	234 (0.53)	30.89

Value in bracket indicating ratio as sum of cement and fly ash equal 1



TABLE II. SIZE DISTRIBUTION OF CRUSHED SAND

Sieve size (mm)	Cumulation retained (g)	Percent retained on each sieve (%)	Cumulative percent retained on sieve Ai (%)
5.00	0	0	
2.50	383.8	38.48	38.48
1.25	225.39	22.60	61.08
0.63	223.96	22.45	83.53
0.315	93.58	9.38	92.92
0.14	31.08	3.12	96.03
< 0.14	39.58	3.97	100.00



(a)



(b)

Figure 4. The fine aggregate investigated: (a) Producing aggregate at An Giang Stone Exploitation & Processing One Member Company [7], (b) Photo of fine aggregate used [3].

### B. Test setup

Experimental tests were conducted using a universal test machine with displacement control, the crosshead speed was set at 1 mm/min during loading. Three specimens for each series were examined then averaged. The universal test machine would record the histories of load and displacement for analysis. The compressive, bending and splitting specimens subject to uniaxial load, three-point bending load and splitting load, respectively.

## III. TESTING RESULTS AND DISCUSSIONS

### A. Indirect tensile strengths

Tables 3 and 4 give the bending strengths and the splitting strengths, respectively, of the tested crushed-sand concretes. Fig. 5 shows the effects of fly ash amount (wt. %) for replacing cement on the failure strengths under compression (a), bending (b) and splitting (c) for both M1 and M2. Regardless of loading types, the maximum strengths was observed at 15 wt.% fly ash replacing cement, and, the strengths of the M2 were always higher than those of the M1.

TABLE III. TENSILE BENDING STRENGTHS OF THE TESTED CRUSHED-SAND CONCRETES

Series	Tensile bending strength (MPa)		
	<i>M1-T00</i>	<i>M1-T15</i>	<i>M1-T30</i>
100x100x300 (M1)	2.30	3.99	2.42
	2.59	3.14	2.65
	3.74	3.10	1.84
Averaged value	2.88	3.41	2.30
Standard deviation	0.76	0.50	0.41
	<i>M2-T00</i>	<i>M2-T15</i>	<i>M2-T30</i>
100x100x300 (M2)	3.64	4.69	4.48
	4.71	4.43	4.78
	4.49	6.51	3.54
Averaged value	4.28	5.21	4.27
Standard deviation	0.57	1.14	0.65

TABLE IV. TENSILE SPLITTING STRENGTHS OF THE TESTED CRUSHED-SAND CONCRETES

Series	Tensile splitting strength (MPa)		
	<i>M1-T00</i>	<i>M1-T15</i>	<i>M1-T30</i>
100x200 (M1)	1.90	1.30	2.50
	2.38	3.00	1.90
	1.70	2.30	1.50
Averaged value	2.00	2.23	1.95
Standard deviation	0.3	0.8	0.5
	<i>M2-T00</i>	<i>M2-T15</i>	<i>M2-T30</i>
100x200 (M2)	2.41	3.18	2.02
	3.91	4.28	3.88
	2.71	3.51	1.86
Averaged value	3.01	3.66	2.59
Standard deviation	0.8	0.6	1.1

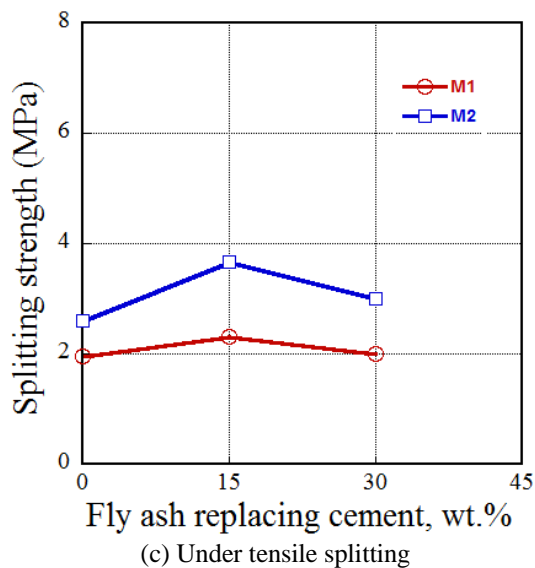
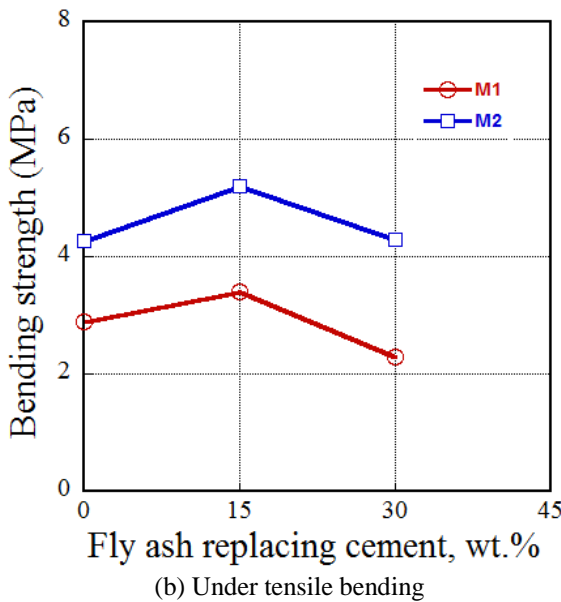
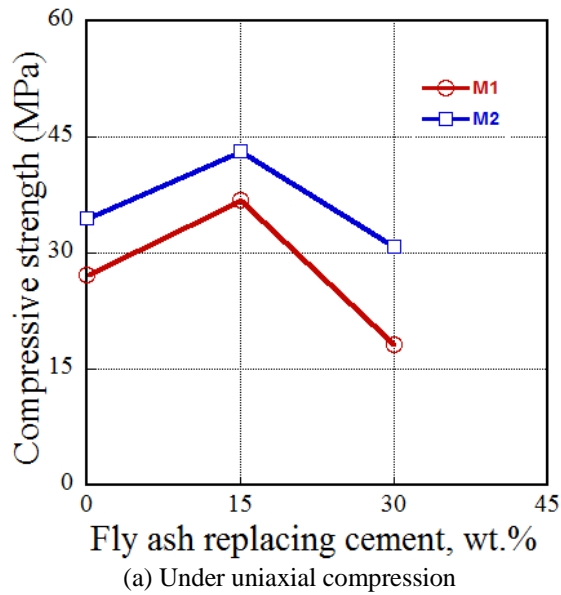


Figure 5. Comparative failure strengths of the crushed-sand concretes regarding fly ash content.

#### B. Scale coefficients in relationship between compressive strength and indirect tensile

To apply the Eqs. 1-3 then to compare with the traditional concrete, the compressive strength of the crushed-sand concrete were firstly translated using conversion factor

$\beta = 1.2$  [8], i.e.,  $f'_c$  in Eqs. 1-3 were derived as follows:

$$f'_c = \sigma_{cylinder}^{150} = \sigma_{cube}^{150} / \beta.$$

Next, the scale coefficients for the bending strengths and splitting strengths of the crushed-sand concrete were computed using Eqs. 2 & 3, respectively. Table 5 shows the values of scale coefficient for bending strengths (value of  $K_{MOR}$  in the brackets) and splitting

strengths ( $K_{SPL}$ ) of the tested crushed-sand concrete. As shown in Table 5, the order of the scale coefficients was same that of traditional concrete, i.e.,  $K_{SPL} < K_{MOR}$  for both M1 and M2 series. Fig. 6 displays the comparisons of the scale coefficients for the traditional concrete and crushed-sand concrete. The  $K_{SPL}$  and  $K_{MOR}$  of the M1 series were smaller than those of traditional concrete (traditional concrete had  $K_{SPL}=0.52$  and  $K_{MOR}=0.63$ ), the differences were 10-28 % for the  $K_{SPL}$  and 2-7 % for the  $K_{MOR}$ . However, the  $K_{SPL}$  and  $K_{MOR}$  of the M2 series were generally higher than those of traditional concrete, the differences were 0-9 % for the  $K_{SPL}$  and 25-38 % for the  $K_{MOR}$ .

TABLE V. SCALE COEFFICIENTS FOR THE INDIRECT TENSILE STRENGTHS OF THE CRUSHED-SAND CONCRETES

Series		Scale coefficient, $K_{SPL}$ ( $K_{MOR}$ )		
		0 wt.% fly ash	15 wt.% fly ash	30 wt.% fly ash
M1	M1-T00	0.42 (0.61)	-	-
	M1-T15	-	0.40 (0.62)	-
	M1-T30	-	-	0.50 (0.59)
		0 wt.% fly ash	15 wt.% fly ash	30 wt.% fly ash
M2	M2-T00	0.56 (0.79)	-	-
	M2-T15	-	0.61 (0.87)	-
	M2-T30	-	-	0.51 (0.84)



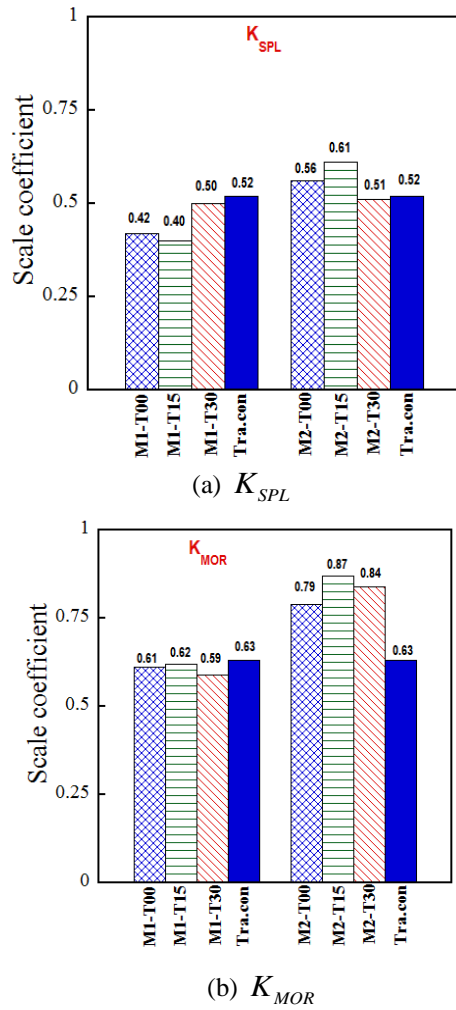


Figure 6. Comparing the scale coefficients between traditional concrete and crushed-sand concrete.

### C. Cracking performance of the crushed-sand concretes



(a) Under tensile bending



(b) Under tensile splitting

Figure 7. Cracking behaviors of the crushed-sand concretes under indirect tensile test.

Fig. 7 presents the photos of typical cracks of the tested specimens at failure. For bending specimen of the crushed-sand concretes, there was a single crack occurring at the middle span. For splitting specimen, there were multiple cracks occurring along the splitting specimen. In both bending and splitting test, the brittle failures were observed with sudden load drop after the peak of the flexural and splitting response curves. In general, the cracking behaviors of the crushed-sand concretes under bending and splitting load was similar to that of traditional concrete.

### IV. CONCLUSIONS

Some conclusions drawn from this research are summarized as follows:

- Indirect tensile responses, including bending response and splitting response, of the crushed-sand concretes, were similar to that of normal concrete. Hence, crushed sand has been possibly used to replace natural river sand in making concrete.
- Fly ash with 15 wt.% of the cement produced the highest indirect tensile strengths for both matrixes M1 and M2. The indirect tensile strengths of the M2 were always higher than those of the M1 and agreed with the tendency of the compressive strength.
- The scale coefficients between indirect tensile strengths and compressive strength of the crushed-sand concrete were as follows: the M1 series produced smaller values and little differences of the scale coefficients in comparison with traditional concrete, while the M2 series produced higher values and much differences.
- To draw the more reliable scale coefficients between indirect tensile strengths and compressive strength of the crushed-sand concrete, further test programs with numerous of specimens are needed.

### ACKNOWLEDGMENT

This research was supported by Ho Chi Minh City University of Technology and Education, the authors are grateful to the sponsor. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsor.

### REFERENCES

- [1] <https://english.vov.vn/society/riverbank-subsidence-scares-mekong-delta-residents-348347.vov>
- [2] <https://english.vietnamnet.vn/fms/environment/162221/mekong-delta-rivers-get-deeper.html>
- [3] Duy-Liem NGUYEN and Minh-Thuan DUONG 2019. Compressive resistance of environmental concrete using fly ash and fine aggregate for replacing traditional sand. Applied Mechanics and Materials, Vol. 889, pp 283-288, doi:10.4028/www.scientific.net/AMM.889.283, © 2019 Trans Tech Publications, Switzerland.
- [4] Balbo J T 2013. Relations between indirect tensile and flexural strengths for dry and plastic concretes. IBRACON Structure and Materials Journal, Volume 6, Number 6 (December 2013) p. 854-874, ISSN 1983-4195.
- [5] American society for testing and materials. Standard Test Method for Flexural Strength of Concrete (using simple beam with third-point loading). ASTM standard C78-08, 2008, West Conshohocken.
- [6] ACI 318-14, Building code requirements for structural concrete.
- [7] <http://khaithacdaangiang.vn/vi/san-pham/dung-lam-be-tong-gach-khong-nung/cat-nghien-28>
- [8] TCVN 8218:2009, Hydraulic concrete - Technical requirements. Vietnamese code.

# A Method to Analyzing and Clustering Aggregate Customer Load Profiles Based on PCA

Alessandro Bosisio  
Energy Department  
Politecnico di Milano  
Milano, Italy  
alessandro.bosisio@polimi.it

Alberto Berizzi  
Energy Department  
Politecnico di Milano  
Milano, Italy  
alberto.berizzi@polimi.it

Andrea Vicario  
Energy Department  
Politecnico di Milano  
Milano, Italy  
andrea.vicario@polimi.it

Andrea Morotti  
Planning Department  
UNARETI S.p.A.  
Milano, Italy  
andrea.morotti@unareti.it

Bartolomeo Greco  
Planning Department  
UNARETI S.p.A.  
Milano, Italy  
bartolomeo.greco@unareti.it

Gaetano Iannarelli  
Astronautics, Electrical and Energetic  
Engineering Department  
Sapienza University  
Rome, Italy  
gaetano.iannarelli@unareti.it

Dinh-Duong Le  
Faculty of Electrical Engineering  
University of Danang-University of  
Science and Technology,  
Danang, Vietnam  
ldduong@dut.udn.vn

**Abstract**— The determination of secondary substations load profiles may facilitate distribution system operators to better forecast, plan and operate their distribution networks. The changing in load and coincidence factors due to the high penetration of distributed generators and loads such as electric vehicles, heating ventilation and air conditioning systems, induction stoves, highlights that standard load profiles are not adapted to the current evolution of the power system. As a result, load profiles analysis has become more significant and valuable. This paper deals with the analysis and clustering of aggregate load profiles by means of principal component analysis. Starting from an extensive field measurement-based database of secondary substations daily load profiles, we used principal component analysis to extract the main components and to reduce data dimension. Thanks to the most significant principal components secondary substations are groups in homogeneous clusters labelled with a standard load profile. The proposed methodology was applied to real load profiles gathered from UNARETI, the distribution system operator of Milano.

**Keywords**—data mining, distribution networks, load profile, pattern clustering, principal component analysis.

## I. INTRODUCTION

Load classification, whether for the system planning and design [1] or the optimized safety operation [2] are extremely important. The current power systems have a wide variety of loads which should have a formal and accurate classification so that similar type of load can be aggregated reducing the difficulty and complexity of load management. Nowadays, the load is growing rapidly, but the analysis of power load characteristics is still at a relatively less studied stage. With the development of economy and society, there are some new types of load in the power grid which may have large differences from the already defined ones. Therefore, it is necessary to reconsider their type and definition.

One of the main issues to be addressed in load classification is that models and data required for the analysis

are dispersed in multiple management systems. The underlying data has not been integrated and unified and analysis cannot be easily performed [3]. Moreover, there are many types of loads in power systems, and such types of load exhibit different characteristics. There are also many kinds of influencing factors for the load characteristics, and the influences for each kind of load are different.

The above problems have restricted the further improvement of the load management and application in power grids led to the difficulties to adapt to the requirements of refined management and technical progress of power grids. Even if load classification is the basis of load forecasting, at present there are many research on power grid load forecasting but relatively few studies on load classification [4]. Therefore, the analysis of the load characteristics and their classification, through appropriate methods, is very important for understanding the changes and trends in electricity grids.

With the development of smart grids, more and more smart meters are installed into distribution networks (DNs) [5], hence, the customers behaviour can be easily and efficiently collected. In this sense, due to the huge amount of data nowadays available, clustering techniques are required for an effective data mining. Moreover, in competitive electricity market with severe uncertainties, performing valuable load classification is important for setting up new tariff offers [6]. The load classification seeks to separate enormous load profiles into several typical clusters [7]. In recent years, researchers have proposed a variety of clustering methods, such as K-means [8]-[9], Fuzzy c-means [10]-[12], hierarchical methods [13]-[14], self-organizing map [15], support vector machine [16]-[17] and subspace projection method [18]. Some papers combining several methods together: in [19], for instance, a method which combines hierarchical and Fuzzy c-mean was introduced, while [20] proposes a two-stage clustering algorithm combined with supervised learning methods to classify electric customers.

With the development of data mining techniques, some new clustering methods have emerged for electricity consumption patterns classification, like in [21], where they built a prediction model to identify the customers who would most likely respond to the prospective offering of the company, or in [22], where the extreme learning machine method is used to analyse the nontechnical loss to classify the customers. Typically, in the data mining, decreasing the dimension of the data and extract the main features of the load profile is an important aspect. In [23], a statistical analysis of end-users historical consumption is conducted to better capture their regularity of consumption, while [24] focusing on the description of the construction and implementation of the recognition of customers risk preference model.

In this paper, a method to analysing and clustering aggregate customer load profiles based on Principal Components Analysis (PCA) is presented. PCA is a very powerful tool that allows the reduction of the order of a system. The technique is based on multivariate statistics analysis that allows to transform a number of possibly correlated variables into a smaller set of variables called Principal Components (PCs). PCA is widely used in many fields of power systems analysis, such as for electromechanical oscillation identification [25] and dynamic thermal rating of transmission lines [26]. Considering a field measurement-based database of Secondary Substations (SSs) daily load profiles, PCA is used to group SSs in homogeneous clusters labelled with a standard load profile. The remaining of this paper is organized as follows: section II explains the theory behind PCA. In section III, the method for load classification is presented. Section IV shows the results of applying the proposed methodology on real load profiles gathered from the management systems of UNARETI, the DSO of Milano. Conclusions are given in section V.

## II. PRINCIPAL COMPONENT ANALYSIS

PCA is a technique exploited in the multivariate statistics that allows to transform a number of possibly correlated variables into a smaller set of variables called Principal Components (PCs) which provide most of the information of the original data [27]. PCA uses a vector space transformation to reduce the dimension of large data sets. Thanks to an orthogonal projection, the original data set can be interpreted in just a few variables, called PCs, so that the component-dependent original variables are converted into new uncorrelated variables. The basic steps of the PCA are as follows:

- Consider the input matrix  $X_{ij}$  of dimension  $n \times m$ , where  $n$  is the number of observations and  $m$  the number of original variables.
- Compute the mean of all rows  $\bar{x}_i$ . The  $\bar{x}_i$  is given by

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$$

and the mean matrix is

$$\bar{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \bar{x}_i$$

- Subtract  $\bar{X}$  from  $X$  to get matrix  $B$ :

$$B = X - \bar{X}$$

- Compute the covariance matrix of the rows of  $B$  by:

$$C = \frac{1}{n-1} B * B$$

- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the PCs. The first PC is  $u_1$  and it is given by

$$u_1 = \operatorname{argmax}_{\|u_1\|=1} u_1 * B * B u_1$$

which is the eigenvector of  $B * B$  corresponding to the largest eigenvalue.  $u_1$  is the left singular vector of  $B$  corresponding to the largest singular value.

As shown in Fig. 1 the new variables, which represent the initial data set in the PCs space, are called *scores*. Scores are constructed as linear combinations of the initial variables weighted by the coefficients of the PCs called *loadings*. These combinations are done in such a way that the new variables are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.

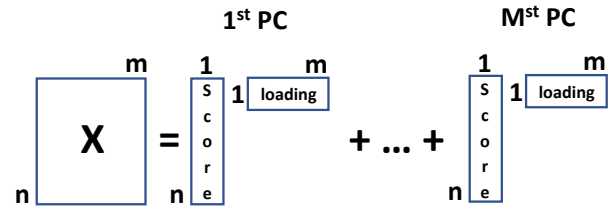


Fig. 1. PCA matrix decomposition.

## III. PROPOSED APPROACH

This section presents the main steps of the proposed load classification approach. As shown in the flow chart of Fig. 2, it consists of the following 4 steps:

- *Applying PCA on input:* PCA is applied on the input matrix  $X_{ij}$  which consists of 60 rows (60 SSs) and 365 columns (365 mean daily power data). PCA returns the PC loadings, scores and their related variance, also known as explained.
- *Selection of the significant PCs:* based on the variance explained by each PC, only the first 3 PCs are considered (PC1, PC2, PC3).
- *Finding correlation of PC loadings with other variables:* for each of the 3 PCs a correlated variable is found in order to give a reasonable meaning based on experience gained from the daily operation of the DN.
- *Classify SSs in clusters:* according to the results of the PCA, SSs are classified in 8 groups. The clusters are defined based on the combination of the positive/negative values of the 3 PC scores.

The procedure finally gives as output several standard aggregate customer load profiles.

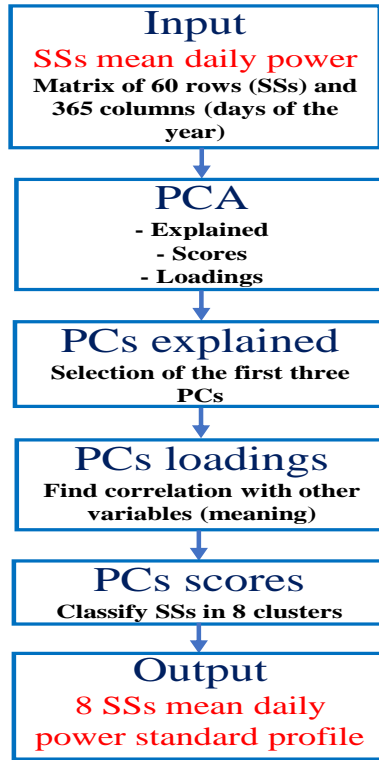


Fig. 2. Flow chart of the proposed approach.

#### IV. RESULTS

The proposed approach has been tested on the 60 SSs shown in Fig. 3. The SSs has been selected to cover the entire metropolitan area of the city of Milano, the service territory of the distribution system operator (DSO) UNARETI, to obtain the maximum variability in terms of load patterns.



Fig. 3. Location of the 60 SSs considered.

The related daily load profiles are shown in Fig. 4. As it is immediately clear, treating the whole data together cannot highlights the main characteristics of the load profiles.

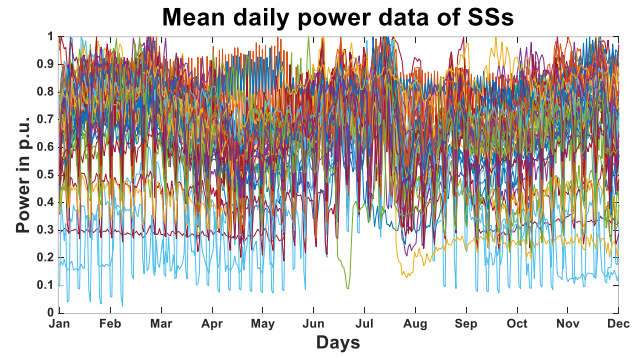


Fig. 4. Mean daily power data of the 60 SSs considered.

##### A. PCA

The input data of the PCA is a 60(SSs)x365(power data for a year) matrix which contains the data shown in Fig. 4. The main output of the PCA is the vector of explained, which consists of 59 rows, the loadings matrix, which has dimension 365x59 and the scores matrix, size 60x59.

##### B. PCs explained

Table I reports the main components extracted by the PCA and the percentage of variance explained by each of them. From the table it is easy to see that the first three PCs can explain almost 97% of the information contain in the input data. Since the goal is to explain most of the information of the original data with the least number of the components, only those three PCs are considered in the following steps.

TABLE I. PCs AND RELATED EXPLAINED

Principal Component	Explained (%)	Accumulated value (%)
1	89.28	89.28
2	5.88	95.17
3	1.43	96.60
4	0.66	97.26
5	0.60	97.86
6	0.34	98.20
7	0.33	98.53
8	0.20	98.72
9	0.14	98.87
10	0.11	98.98
>10	1.02	100

##### C. PCs loadings

The following step of the procedure analyses the PCs loadings in order to find a feasible physical meaning for each of them. This is done by finding correlated variables to the PCs based on the investigation and knowledge of the phenomena related to power systems.

##### a) First Principal Component (PC1)

Fig. 5 shows, in the upper graph, the loadings of the PC1 and in the lower graph the daily load profile of the DN of Milano. The correlation of the two curves is almost 90% which can suggest that the PC1 counts the variability of the load demand during the year. The load profile of the Milano DN, as well as loadings of PC1, has valleys in holiday periods, i.e. Christmas and New Year's Eve, Easter and Summer between July and August. Hence, since loadings of PC1 are higher in working days, PC1 may represent the activities factor.



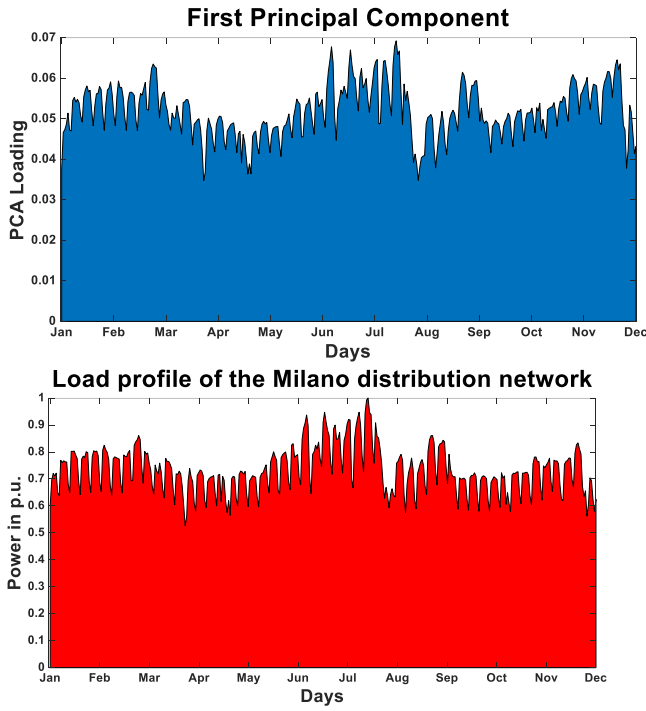


Fig. 5. Loadings of the first PC (first chart) Vs load profile of the Milano distribution network (second chart).

In Fig. 6 the yearly load profile of the SSs with the highest and the lowest PC1 are shown. The SS with the highest PC1 has a similar trend to the load profile of the Milano DN while the other SS, the one with the lowest PC1, does not follow this evolution and has a more uniform profile over the year.

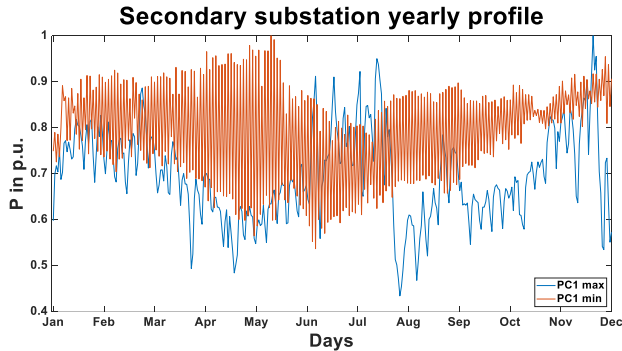


Fig. 6. Yearly profile of the SS with the highest PC1 Vs yearly profile of the SS with the lowest PC1.

#### b) Second Principal Component (PC2)

Fig. 7 shows, in the upper graph, the loadings of the PC2 and in the lower graph the daily temperature recorded in Milano. The correlation of the two curves is almost 85% which can suggest that the PC2 weights the influence of hot weather on the SSs load profile. In fact, the loadings are positive during the summer season and negative in the others. Hence, PC2 may represent the air conditioning factor, i.e. the use of air conditioners by customers.

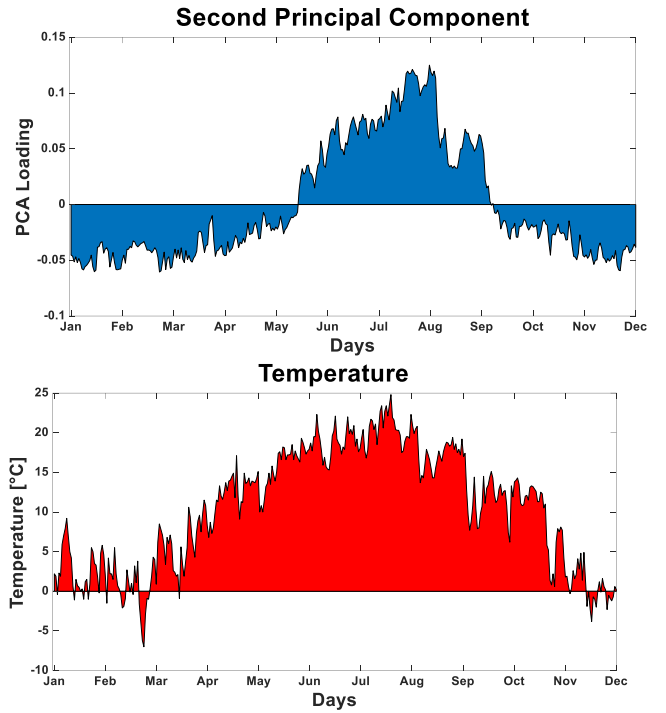


Fig. 7. Loadings of the second PC (first chart) Vs mean daily temperature recorded in Milano (second chart).

In Fig. 8 the yearly load profile of the SSs with the highest and the lowest PC2 are depicted. On the one hand, the SS with the highest PC2 has a significant increase in power demand between middle May and September. On the other hand, the SS with the lowest PC2 decreases the power demand in summer months.

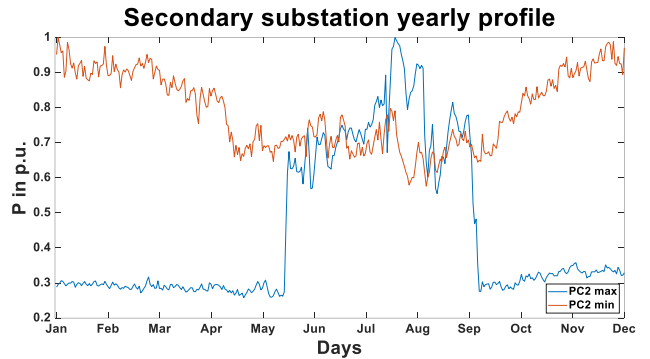


Fig. 8. Yearly profile of the SS with the highest PC2 Vs yearly profile of the SS with the lowest PC2.

#### c) Third Principal Component (PC3)

Fig. 9 shows the loadings of the PC3. Having a look to Fig. 10, it is immediately clear that the PC3 counts the variability of the load during the week. In fact, the cycle of 5 weekdays and 2 weekend days can be found. PC3 considers holidays, i.e. Christmas and New Year's Eve, Easter and Summer between July and August as weekend days. This can be explained considering the typical decreasing in load demand in those periods.

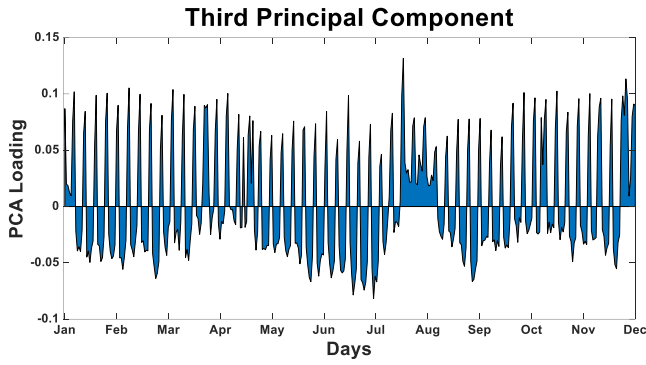


Fig. 9. Loadings of the third PC.

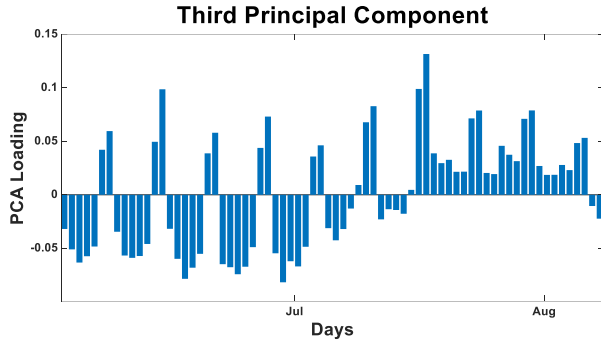


Fig. 10. Loadings of the third PC (zoom in).

In Fig. 11 the yearly load profile of the SSs with the highest and the lowest PC3 are shown. The SS with the lowest PC3 has a trend which follows the difference between weekdays and weekend while the one with the highest PC3, as shown also in Fig. 12, has a more uniform shape through the weekdays.

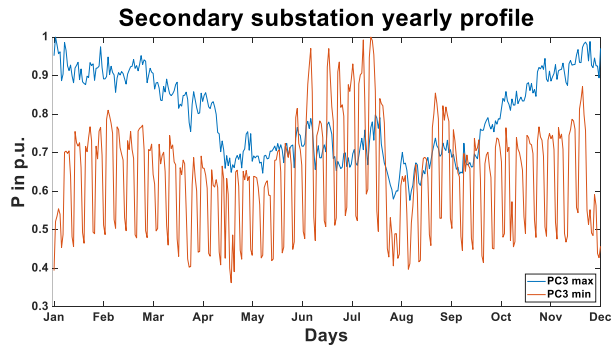


Fig. 11. Yearly profile of the SS with the highest PC3 Vs yearly profile of the SS with the lowest PC3.

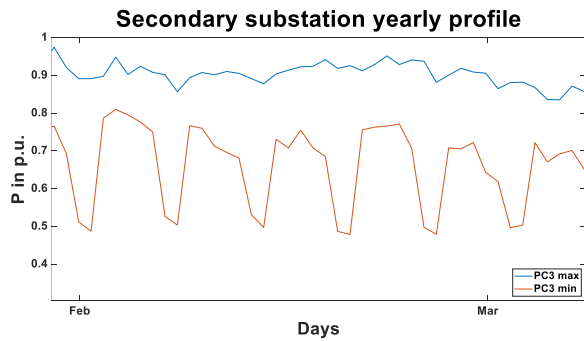


Fig. 12. Yearly profile of the SS with the highest PC3 Vs yearly profile of the SS with the lowest PC3 (zoom in).

#### D. PCs scores

The last step of the proposed methodology analyses the PCs scores to group similar SSs in clusters. Fig. 13 shows the 2-D scatter plot of the PC1, on the X-axis, and on the PC2 on the Y-axis. As expected, PC1 express a wider variability than PC2. Fig. 14 shows the relationship between PC1 and PC3 while Fig. 15 between PC2 and PC3. The two figures confirm that PCA shares a decreasing variability while going down to lower order components.

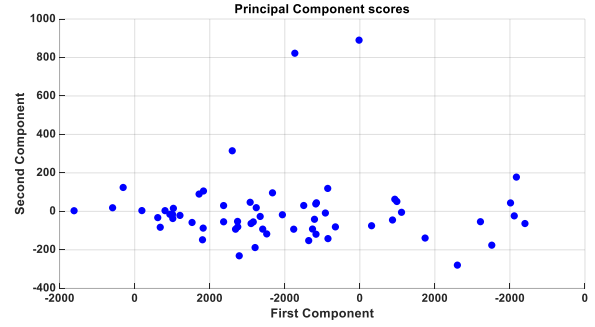


Fig. 13. 2-D scatter plot of the first and the second PCs.

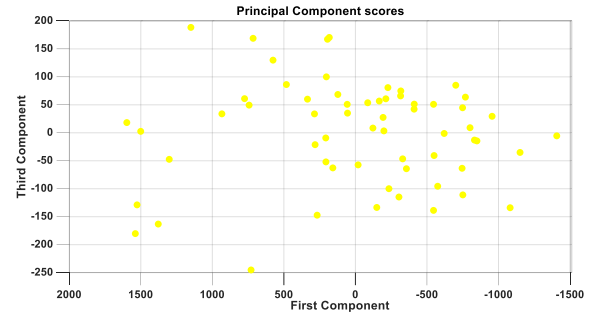


Fig. 14 2-D scatter plot of the first and the third PCs

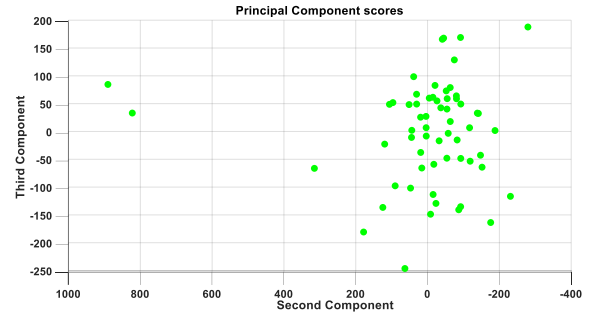


Fig. 15. 2-D scatter plot of the second and the third PCs.

Based on the 3-D scatter plot reported in Fig. 16, the 60 SSs are classified using the cube shown in Fig. 17.

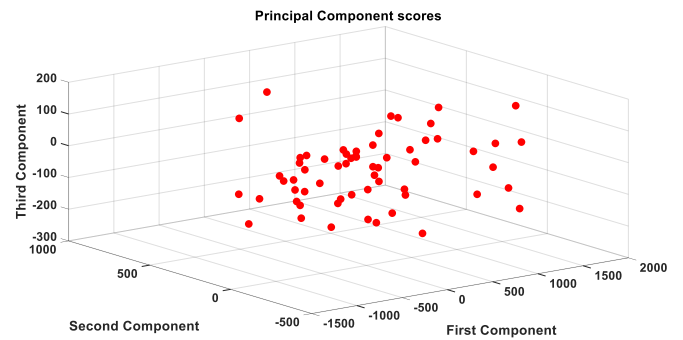


Fig. 16. 3-D scatter plot of the first three PCs.

Since we considered only the first three components, a reasonable way to classify SSs it should be therefore based on the positive/negative combination of their scores. For instance, a SS which has all the PC scores positive is included in the sub cube 1; a SS with negative PC1, positive PC2 and PC3 is included in the sub cube VIII. Following this approach, each SS is located, i.e. clustered, in one of the 8 sub cubes.

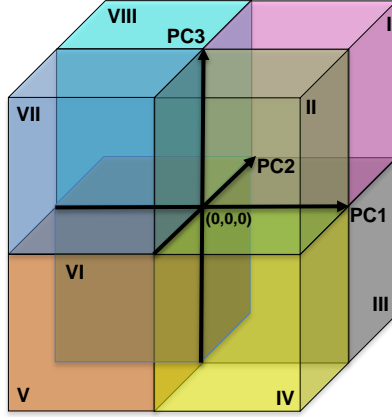


Fig. 17. 3-D Cube used for SSs classification.

Table II reports the PCs sign for each of the 8 clusters and also gives a definition of the related load profile based on the outcome of loadings analysis. For instance, a SS which fall in the cluster 1 has a load profile which is correlated to the load profile of the Milano DN, to the hot weather and to the cycle weekdays/weekend. The last column of the table reports the number of SS in each cluster. The most common types are number II, V and VII.

TABLE II. CLUSTERS AND THEIR MAIN CHARACTERISTICS

Cluster	(PC1,PC2,PC3)	Main characteristics of the load profile	# SS
I	(+,+,+)	DN load profile correlated, hot weather correlated, weekdays/weekend uncorrelated	6
II	(+,-,+)	DN load profile correlated, hot weather uncorrelated, weekdays/weekend uncorrelated	11
III	(+,+,-)	DN load profile correlated, hot weather correlated, weekdays/weekend correlated	4
IV	(+,-,-)	DN load profile correlated, hot weather uncorrelated, weekdays/weekend correlated	6
V	(-,-,-)	DN load profile uncorrelated, hot weather uncorrelated, weekdays/weekend correlated	10
VI	(-,-,+)	DN load profile uncorrelated, hot weather correlated, weekdays/weekend correlated	7
VII	(-,+,-)	DN load profile uncorrelated, hot weather uncorrelated, weekdays/weekend uncorrelated	11
VIII	(-,+,-)	DN load profile uncorrelated, hot weather correlated, weekdays/weekend uncorrelated	5

Once the SSs are all associated to a group, it is also possible to extract a standard load profile for each cluster, simply average the SS load profiles day by day. Fig. 18 shows the load profiles of the SSs which belong to cluster IV (+,-,-).

Looking at the standard profile of Fig. 19, it is possible to recognise the high correlation with the load profile of the Milano DN, the low correlation with hot weather and the high correlation with the cycle weekdays/weekend.

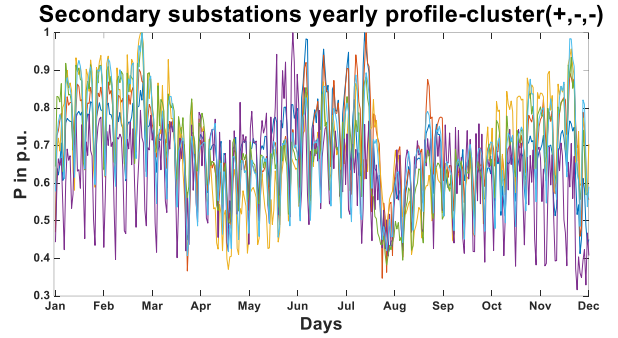


Fig. 18. Load profile of SSs classified in cluster IV (+,-,-).

Secondary substation yearly standard profile-cluster(+,-,-)

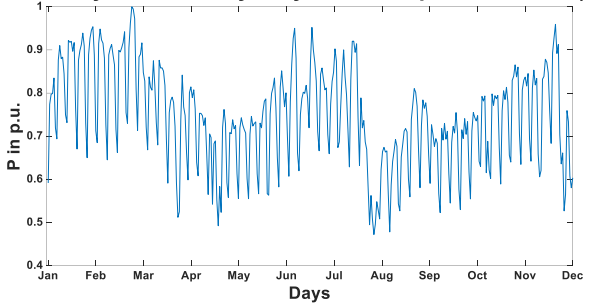


Fig. 19. Standard load profile for a SS classified in cluster IV (+,-,-).

Fig. 20 reports, instead, the load profiles of the SSs which belong to cluster VI (-,-,+). Looking at the standard profile of Fig. 21, it is possible to recognise the low correlation with the load profile of the Milano DN, the high correlation with hot weather (the power demand strongly increases in summer) as well as the high correlation with the cycle weekdays/weekend.

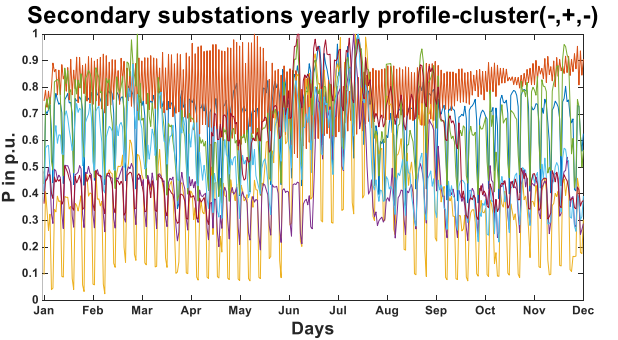


Fig. 20. Load profile of SSs classified in cluster VI (-,-,+).

Secondary substation yearly standard profile-cluster(-,-,+)

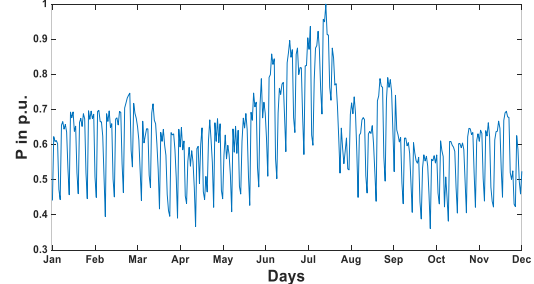


Fig. 21. Standard load profile for a SS classified in cluster VI (-,-,+).



Finally, Fig. 22 depicts the load profiles of the SSs belong to cluster VIII (-,+,+). Looking at the standard profile of Fig. 23, it is possible to recognise the low correlation with the load profile of the Milano DN, the high correlation with hot weather and the low correlation with the cycle weekdays/weekend (the power demand does not always follow the passing of the week days).

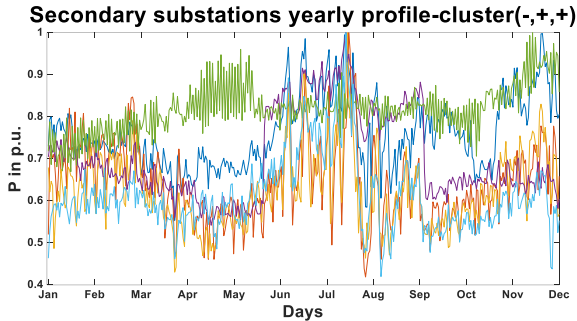


Fig. 22. Load profile of SSs classified in cluster VIII (-,+,+).

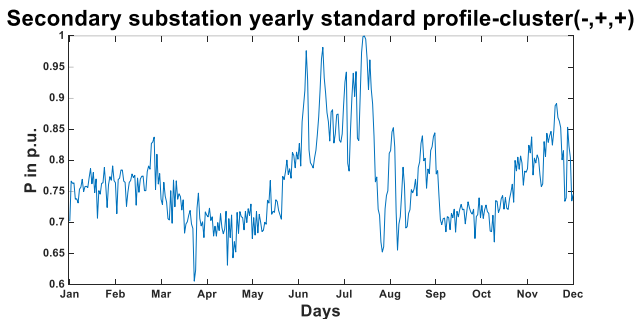


Fig. 23. Standard load profile for a SS classified in cluster VIII (-,+,+).

## V. CONCLUSIONS

In this paper, real daily power data of several SSs were analysed, by means of PCA, to get standard profiles. The first three main components of the PCA have been shown to represent more than 96% of the total variance of the input data and to be correlated with the Milano DN load curves, the temperature and the cycle of weekdays and holidays respectively. The SS load profiles have been then clustered in 8 groups with different characteristics related to the positive/negative values of the PC scores. The standard load profiles identified by the proposed methodology will give to the DSO of Milano very useful and valuable information for better planning, operation, and forecasting its distribution network.

## REFERENCES

- [1] A. Bosisio, A. Berizzi, C. Bovo, E. Amaldi, and S. Fratti, "GIS-based urban distribution networks planning with 2-step ladder topology considering electric power cable joints," in *2018 110th AEIT International Annual Conference, AEIT 2018*, 2018.
- [2] A. Bosisio, A. Berizzi, A. Morotti, A. Pegoiani, B. Greco, and G. Iannarelli, "IEC 61850-based smart automation system logic to improve reliability indices in distribution networks," in *2019 AEIT International Annual Conference, AEIT 2019*, 2019.
- [3] A. Bosisio, D. D. Giustina, S. Fratti, A. Dede, and S. Gozzi, "A metamodel for multi-utilities asset management," in *2019 IEEE Milan PowerTech, PowerTech 2019*, 2019.
- [4] C. Kuster, Y. Rezgui, and M. Mourshed, "Electrical load forecasting models: A critical systematic review," *Sustainable Cities and Society*, vol. 35. Elsevier Ltd, pp. 257–270, 01-Nov-2017.
- [5] Q. LI, Z. XU, and L. YANG, "Recent advancements on the development of microgrids," *J. Mod. Power Syst. Clean Energy*,

- vol. 2, no. 3, pp. 206–211, Jan. 2014.
- [6] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 381–387, 2003.
- [7] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, "Clustering load profiles for demand response applications," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1599–1607, 2019.
- [8] G. J. Tsekouras, N. D. Hatziaargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, 2007.
- [9] N. M. Kohan, M. P. Moghaddam, S. Mohammad Bidaki, and G. R. Yousefi, "Comparison of Modified K-Means and Hierarchical Algorithms in Customers Load Curves Clustering for Designing Suitable Tariffs in Electricity Market," *Proc. Univ. Power Eng. Conf.*, pp. 1–5, 2008.
- [10] Z. Zakaria, K. L. Lo, and M. H. Sohod, "Application of fuzzy clustering to determine electricity consumers' load profiles," *First Int. Power Energy Conf. (PECon 2006) Proc.*, pp. 99–103, 2006.
- [11] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Jarventausta, "Enhanced load profiling for residential network customers," *IEEE Trans. Power Deliv.*, vol. 29, no. 1, pp. 88–96, 2014.
- [12] S. Tao, Y. Li, X. Xiao, and L. Yao, "Load forecasting based on short-term correlation clustering," *2017 IEEE Innov. Smart Grid Technol. - Asia Smart Grid Smart Community, ISGT-Asia 2017*, pp. 1–7, 2018.
- [13] G. Chicco, R. Napoli, F. Piglion, P. Postolache, M. Scutariu, and C. Toader, "Emergent electricity customer classification," *IEE Proc. Gener. Transm. Distrib.*, vol. 152, no. 2, pp. 164–172, 2005.
- [14] G. Chicco, R. Napoli, and F. Piglion, "Comparisons Among Clustering Techniques for Electricity Customer Classification," vol. 21, no. 2, pp. 933–940, 2006.
- [15] S. V. Verdú, M. O. García, C. Senabre, A. G. Marín, and F. J. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1672–1682, 2006.
- [16] G. Chicco and I. S. Ilie, "Support vector clustering of electrical load pattern data," *IEEE Trans. Power Syst.*, vol. 24, no. 3, pp. 1619–1628, 2009.
- [17] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Deliv.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.
- [18] M. Piao, H. S. Shon, J. Y. Lee, and K. H. Ryu, "Subspace projection method based clustering analysis in load profiling," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2628–2635, 2014.
- [19] X. Lin, W. Wu, B. Zeng, X. Yan, S. Han, and L. Qin, "Analysis of large-scale electricity load profile using clustering method," *ICNSC 2018 - 15th IEEE Int. Conf. Networking, Sens. Control*, pp. 1–5, 2018.
- [20] B. Peng et al., "A two-stage pattern recognition method for electric customer classification in smart grid," *2016 IEEE Int. Conf. Smart Grid Commun. SmartGridComm 2016*, pp. 758–763, 2016.
- [21] T. K. Das, "A customer classification prediction model based on machine learning techniques," *Proc. 2015 Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2015*, pp. 321–326, 2016.
- [22] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, 2008.
- [23] J. Yang, J. Zhao, F. Wen, and Z. Dong, "A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3374–3386, 2019.
- [24] B. Liu, H. Qiu, and Y. Shen, "Establishment and implementation of securities company customer classification model based on clustering analysis and PCA," *Proc. - 2012 Int. Conf. Control Eng. Commun. Technol. ICCECT 2012*, pp. 325–329, 2012.
- [25] A. Bosisio et al., "Combined use of PCA and Prony Analysis for Electromechanical Oscillation Identification," in *ICCEP 2019 - 7th International Conference on Clean Electrical Power: Renewable Energy Resources Impact*, 2019, pp. 62–70.
- [26] A. Bosisio, A. Berizzi, D.-D. Le, F. Bassi, and G. Giannuzzi, "Improving DTR assessment by means of PCA applied to wind data," *Electr. Power Syst. Res.*, vol. 172, pp. 193–200, Jul. 2019.
- [27] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering*. Cambridge University Press, 2019.

# The Bearing Capacity of Compacted Clay Reinforced by Geotextile and Sand Cushion

Minh-Duc Nguyen

Faculty of Civil Engineering

Ho Chi Minh City University of Technology and Education

Ho Chi Minh City, Vietnam

ducnm@hcmute.edu.vn

**Abstract**— The paper presents a series of laboratory tests for the California Bearing Ratio (CBR) for investigating the bearing capacity of soft clay reinforced by nonwoven geotextile and sand cushion. The variation of test conditions included levels of compaction energy and the thickness of sand layers. The result reveals that a thin sand cushion layer and reinforcement inclusion improved the CBR value of the reinforced riverbed clay significantly. The optimum thickness of sand cushion was 1.5 cm, which equivalent to the ratio between the depth of the first reinforcement layer and the diameter of the penetrated piston is equal unity. The CBR of the reinforced specimens could be increased as high as 25.5-34.8% at this optimum reinforcement ratio.

**Keywords**— CBR, nonwoven geotextile, soft clay, sand cushion

## I. INTRODUCTION

In recent years, the development of infrastructure in the Mekong Delta area in Vietnam significantly increased the needs of sandy soil as the backfill material. However, as the unrecoverable material, the sand was excavated from the Mekong River, which would have several adverse effects for society, including high construction cost, riverside instability, and erosion. The soft clay as local material was preferred for rural road construction; however, it requires further bearing capacity improvement before applying on the construction.

There were various methods to improve the strength behavior of the soft clay, including the geosynthetic with a thin granular soil layer. The improvement of strength and deformation behavior of the reinforced soft clay with a thin sandy soil layer were investigated using direct shear tests [1], pullout tests [2-4], and triaxial compression tests [5-6]. The results revealed that the shear strength of the reinforced clay was significantly improved due to the increase in the interface interaction between the clay, the reinforcement, and the thin sand-layer inclusion. The anisotropic shear strength behavior of reinforced soil showed that the geosynthetic reinforcement should be placed horizontally to enhance the confinement caused by the lateral tensile strain mobilized in the soil transferred to the geosynthetic layers [7]. The excess pore water pressure during shearing in the reinforced clay was dissipated laterally to the sandy layer [8]. Besides, the optimum thickness of the granular soil layer for the highest shear strength improvement was inconsistently concluded in the previous studies. The optimal thickness of the thin sand layer was observed in [1, 3-4].

On the other hand, under undrained, unconsolidated triaxial compression test, the shear strength improvement increased with sand thickness of 5-20mm [5]. In those studies, the reinforcement layer was embedded in a thin sand layer

(sandwich technique) without separation between the soft clay and sand layers. Due to the seepage process, the clay particles gradually clogged into the thin sandy soil layer and reduced its permeability. It would decrease the effect of the thin sandy layer on dissipating water pressure in the reinforced clay. Besides, the thin sand layer was vulnerable under the erosion effects of the seepage.

Several previous studies used the laboratory test for the California Bearing Ratio (CBR) to investigate the bearing capacity of reinforced clay [9-16]. The laboratory results showed that the CBR of the geosynthetic clay liners (GCLs) under a sand layer reached the most significant improvement when the thickness of the sand layer equal to the diameter of the load piston [9]. On the other hand, [10-12] concluded that placing a geogrid layer in the middle of the base layer effectively reduced the settlement of reinforced pavement. A similar observation was also found for the clay reinforced by a single jute textile layer [13]. On the other hand, the optimum location of a single geogrid layer was observed to be taken as 72-76% of specimen height for clean sand (SP), sandy clay (CL), and clayey silt (ML) [14].

Regarding reinforced clay, few studies have focused on the effects of a thin sand layer inclusion on improving its bearing capacity. Accordingly, this study conducted a series of laboratory tests for CBR values on the clay specimens reinforced by a thin sand layer covered with two nonwoven geotextile layers. The proposed reinforcement arrangement was expected to prevent the sand erosion and clay-sand clogging due to the seepage in the reinforced soil. This study's main objective was to investigate the bearing capacity behavior of reinforced clay with a thin sand cushion layer. The study results provided useful information for effectively improved the bearing capacity of the reinforced soft clay using a new arrangement of reinforcement inclusion with a thin sand layer.

## II. TEST MATERIALS

### A. Soft clay

The soft clay was excavated from Cai Lon, Kiengiang river in the Mekong Delta, Vietnam. It is classified as high plastic inorganic silt (MH) by the Unified Soil Classification System (Fig. 1). Its liquid limit (LL), plastic limit (PL), and specific gravity are 91.5, 44.9, and 2.75, respectively. Table 1 shows the results of the modified Proctor test, in which the clay was compacted in 5 layers using the modified rammer. The value of optimum moisture content, OMC decreased when the compaction energy increased from 482 kJ/m<sup>3</sup> to 2700 kJ/m<sup>3</sup>.

TABLE I. MODIFIED PROCTOR TEST RESULTS

Compaction energy, $E$ (kJ/m <sup>3</sup> )	Total number of blows	OMC (%)	Maximum dry unit weight (kN/m <sup>3</sup> )
482	50	26.6	14.28
1200	125	24.5	15.02
2700	280	20.5	16.31

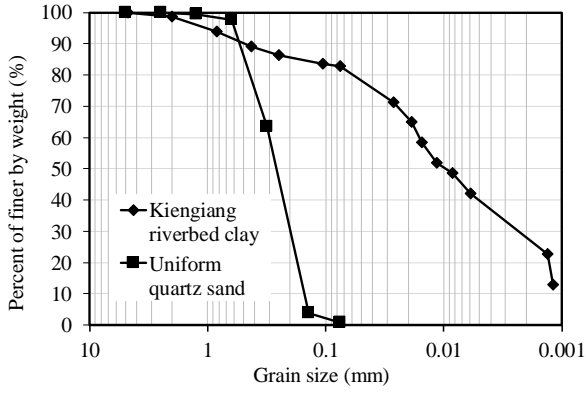


Fig. 1. Grain size distribution of the Kiengiang riverbed clay and the uniform quartz sand

### B. Uniform quart sand

The local sand is uniform and clean quartz sand, classified as poorly graded sand (SP) by the Unified Soil Classification System. The specific gravity ( $G_s$ ) coefficient of uniformity ( $C_u$ ), and gradation ( $C_c$ ) were 2.66, 2.00, and 0.98, respectively (Fig. 1). The minimum and maximum dry unit weights of sand were  $\gamma_{d-\min} = 12.56$  kN/m<sup>3</sup> and  $\gamma_{d-\max} = 12.53$  kN/m<sup>3</sup>. At 90% of relative density, the shear strength parameters were obtained from the direct shear test as  $c = 0$  and  $\phi = 35.1^\circ$ . The sand-geotextile interface friction angle was  $\phi'_a = 23.7^\circ$ , measured by a modified direct shear test (the top shear box was filled with soil, and steel platen was placed in the lower one).

### C. Nonwoven geotextile

A commercially PET nonwoven geotextile has a permittivity of  $\gamma = 1.96$  s<sup>-1</sup> and corresponding cross-plane permeability of  $k = 3.5 \times 10^{-3}$  m/s. Table 2 shows its ultimate tensile strength and failure strain along the longitudinal and transverse directions measured from the wide-width tensile test. The apparent opening size,  $O_{90}$  is about 0.11 mm, which is smaller than the diameter at 10% of percent passing,  $D_{10}$  of the uniform sand, which ensuring the proper filter function of the geotextile layer to separate the soft clay and the sandy soil in the reinforced specimens.

TABLE II. PROPERTIES OF GEOTEXTILE

Property	Value	
Mass (g/m <sup>2</sup> )	200	
Thickness (mm)	1.78	
Apparent opening size (mm)	0.11	
Permittivity (s <sup>-1</sup> )	1.96	
Cross-plane permeability (m/s)	3.5×10-3	
Wide-width tensile test		
Direction	Ultimate strength (kN/m)	Failure strain (%)
Longitudinal	9.28	84.1
Transverse	7.08	117.8

## III. EXPERIMENTAL PROGRAM

A total of 15 laboratory tests was conducted to determine the CBR value of the reinforced clay with sand cushion. The test variation included the thickness of sand cushion layer and compaction energy levels.

### A. Specimen preparation

After being excavated from the riverbed, the soft clay was dried in an oven (less than 60°C) for about 24h then crushed and ground into powder in a mortar. The dried powder clay then mixed with water corresponding to the desired optimum moisture content (Table 1), stored in a resalable plastic in a plastic bag within a temperature-controlled chamber for a minimum of 2 days to ensure a uniform distribution of moisture within the soil mass.

The specimens for CBR test were compacted in a mold with a diameter,  $D = 152.4$  mm, and a height,  $H = 116.6$  mm. The samples were statically compressed to the required thickness by a static hydraulic jack. To prepare the reinforced clay specimens, firstly, the lower mold was filled with clay in several layers. The amount of soil of each layer corresponded to the soil density obtained from compaction test results. After compacting a clay layer, it was scarified prior to pouring the next soil layer until reaching the desired thickness of the lower clay layer. Its surface was also scarified then added a geotextile layer for developing good interface bonding between the clay and the reinforcement layer. The sand cushion layer was then added and compacted by a tamper to reach 90% of relative density. After that, another geotextile layer was placed above the sand layer after compaction of the sandy soil layer. The upper clay layer was finally compacted to complete the reinforced specimens (Fig. 2).

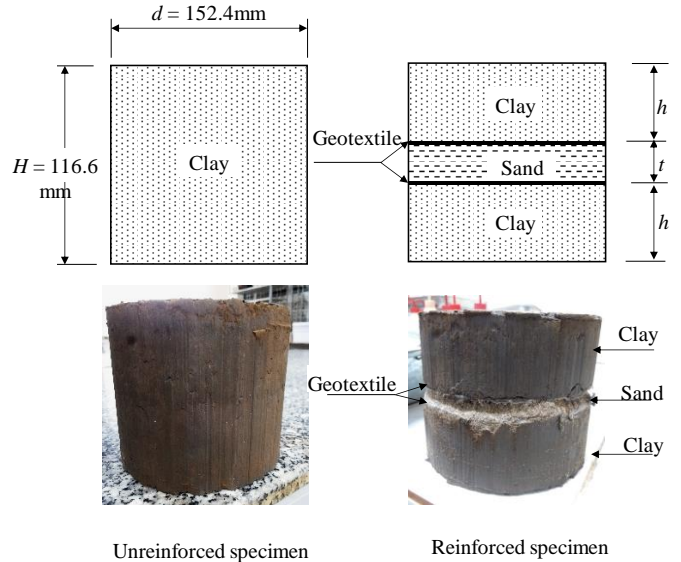


Fig. 2. The reinforced and unreinforced specimens in laboratory test for CBR values

### B. Testing program

The laboratory test for the CBR value was performed following [17]. A mass of surcharge (i.e., 4.54 kg) was placed on the specimens before applying the load on a piston with a diameter,  $B = 49.7$  mm, penetrated the specimens. The rate of penetration was approximately 0.05 inch/min (1.27 mm/min). The tests were stopped until the penetration reached 20 mm. During the penetration of the piston, the compression force was recorded with time. It was then corrected due to the

surface irregularities or other causes, as recommended by [17], before evaluating the CBR value from  $CBR_1$  and  $CBR_2$ .

$$CBR_1(\%) = \frac{P_1}{6900} \times 100 \quad (1)$$

$$CBR_2(\%) = \frac{P_2}{10300} \times 100 \quad (2)$$

where  $P_1$  and  $P_2$  are the corrected stresses in kPa at 2.54 mm and 5.08 mm of penetration, respectively. The CBR value is chosen as the higher value of  $CBR_1$  and  $CBR_2$ . In general,  $CBR_1$  is higher than  $CBR_2$ , and  $CBR$  is equal to  $CBR_1$ . If  $CBR_2 > CBR_1$ ,  $CBR_2$  is chosen as the CBR value after the redo tests to confirm the accuracy of the original test result [17].

#### IV. RESULTS AND DISCUSSION

##### A. CBR behavior of unreinforced and reinforced specimens

Figures 3 shows the variation of the corrected stress in the penetration piston of unreinforced and reinforced specimens with the penetration. The bearing capacity of clay specimens was significantly improved due to the reinforcement inclusion and the sand cushion layer. Similar conclusions on the bearing capacity improvement of the reinforced clay were reported in numerous studies [8-15, 18].

The CBR values of the unreinforced and reinforced evaluated from the corrected piston stress were shown in Table 3. The results show that the higher the compaction energy was, the higher CBR value of specimens was obtained. For all the specimens,  $CBR_2$  was lower than  $CBR_1$ , which was chosen as the CBR value of the specimens [17].

Compared to the bearing capacity of unreinforced specimens, the CBR value of reinforced specimens was significantly improved. The bearing capacity of reinforced specimens depended on the depth of the top reinforcement layer,  $h$  (Fig. 2). For the three compaction energy levels, when increasing the thickness of the sand cushion layer, the CBR value initial increased from that of the unreinforced specimen,  $h=11.66$ cm (i.e.,  $l=0$  cm) reached a peak value for the reinforced specimen with 1.5cm of the thickness of sand cushion ( $h=4.93$ cm) then reduced. Thus, the optimum value,  $h/B=1.0$ , was found for maximum bearing capacity of the specimens reinforced by geotextile, and the sand cushion layer was. It was in agreement with those reported in the previous studies [8, 10-11]. On the other hand, [13] reported that a geogrid layer was placed at a depth of about 1.0-1.2 times the diameter of the plate load to attain the highest CBR value of reinforced specimens.

The bearing capacity improvement due to reinforcement inclusion and the sand cushion layer was quantified using the bearing capacity difference and percent CBR improvement,  $\Delta CBR$ . The bearing capacity difference was defined as the difference corrected stress of the reinforced and unreinforced specimens at the same penetration. While the parameter  $\Delta CBR$  was evaluated as:

$$\Delta CBR = \frac{CBR_{re} - CBR_{unre}}{CBR_{unre}} \times 100\% \quad (3)$$

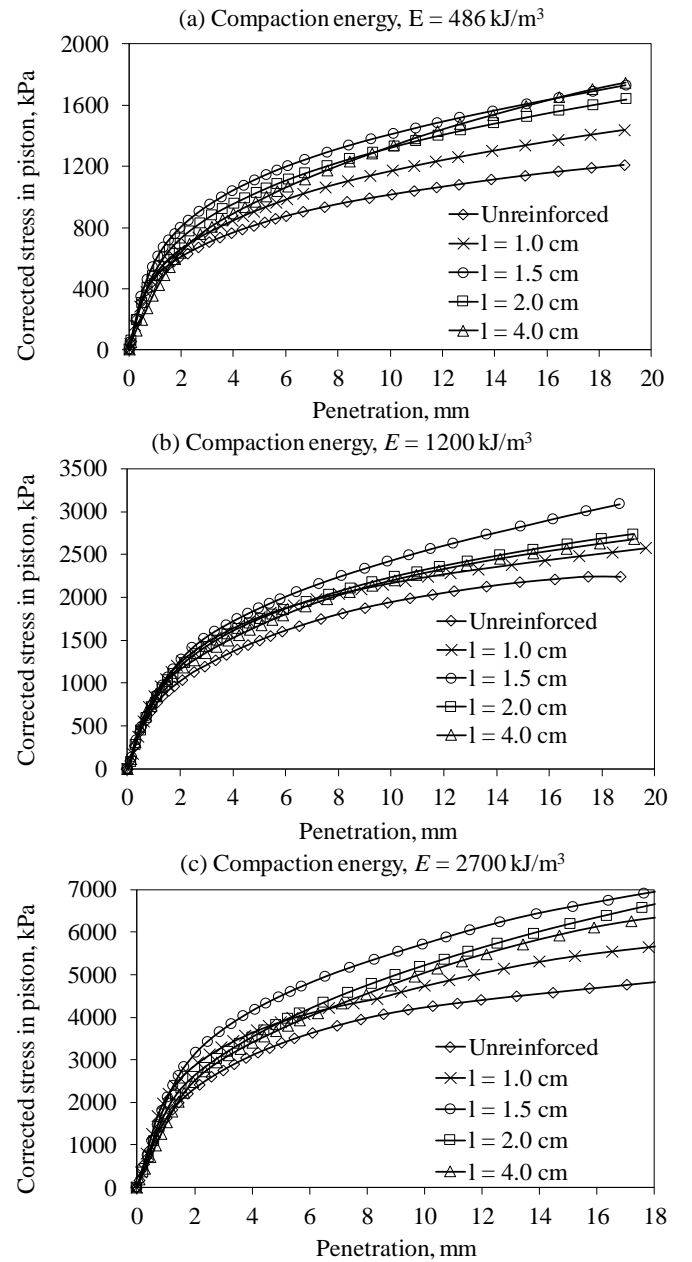


Fig. 3. Corrected stress in piston of unreinforced and reinforced specimens compacted by 3 different compaction energy levels: (a) 482 kJ/m<sup>3</sup>; (b) 1200 kJ/m<sup>3</sup>; (c) 2700 kJ/m<sup>3</sup>

As shown in Figure 4, the increment of piston penetration induced more bearing capacity difference in the reinforced specimens. It can be explained as the more piston penetration induced higher tensile strength mobilized in the geotextile layers to improve the bearing capacity of the reinforced specimens, which was reported in the previous researches.

[19] performed the consolidated drained triaxial compression test to investigate the shear strength of the reinforced sand. It was observed that the mobilized shear strength of reinforced soil exceeded that of unreinforced soil under a range of axial strain of approximately 1–3%.

TABLE III. RESULTS OF UNREINFORCED AND REINFORCED SPECIMENS

Compaction energy, E (kJ/m <sup>3</sup> )	Thickness of sand cushion layer, <i>l</i> (cm)	<i>h/B</i>	CBR <sub>1</sub> (%)	CBR <sub>2</sub> (%)	CBR (%)
482	0 (unreinforced)	2.4	9.6	8.0	9.6
482	1	1.1	10.4	9.0	10.4
482	1.5	1.0	12.8	11.1	12.8
482	2	0.9	11.7	10.1	11.7
482	4	0.7	10.6	9.6	10.6
1200	0 (unreinforced)	2.4	16.5	14.6	16.5
1200	1	1.1	19.9	17.3	19.9
1200	1.5	1.0	20.7	18.3	20.7
1200	2	0.9	19.3	17.0	19.3
1200	4	0.7	18.4	16.3	18.4
2700	0 (unreinforced)	2.4	37.4	33.0	37.4
2700	1	1.1	45.3	38.0	45.3
2700	1.5	1.0	50.4	43.9	50.4
2700	2	0.9	42.3	38.0	42.3
2700	4	0.7	40.8	36.4	40.8

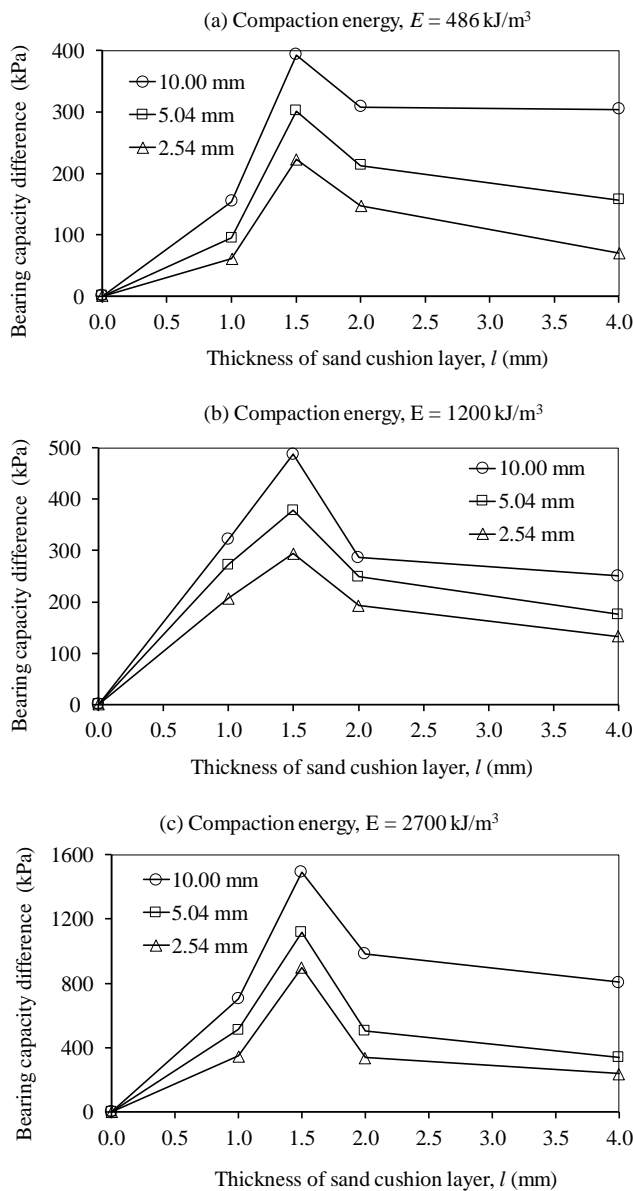


Fig 4. The variation of bearing capacity difference at different piston penetration 2.54mm; 5.08mm and 10.00mm.

Similar to the variation of CBR values on the thickness of the sand layer, the bearing capacity difference also peaked at the specimens reinforced with 1.5cm thickness of the sand cushion layer, which was equivalent to the optimum ratio,  $h/B = 1$ .

The influence of compaction energy on the bearing capacity difference was shown in Fig. 5. It was observed that increment in compaction energy produced more significant bearing capacity improvement of the reinforced specimens. The more compaction energy, the higher the interface shear strength of clay and geotextile layers was. As a result, it induced more mobilized tensile strength in the geotextile to enhance the bearing capacity of the reinforced specimens. The results demonstrate the benefit of using the high compaction energy for compacting the clay layers in the reinforced specimens.

On the other hand, the percent CBR enhancement,  $\Delta\text{CBR}$  of the reinforced clay changed inconsistently with the changes of compaction energy. For the specimens reinforced by 1.5cm thickness of the sand cushion layer, the value of  $\Delta\text{CBR}$  could be up to 25.5-34.8%.

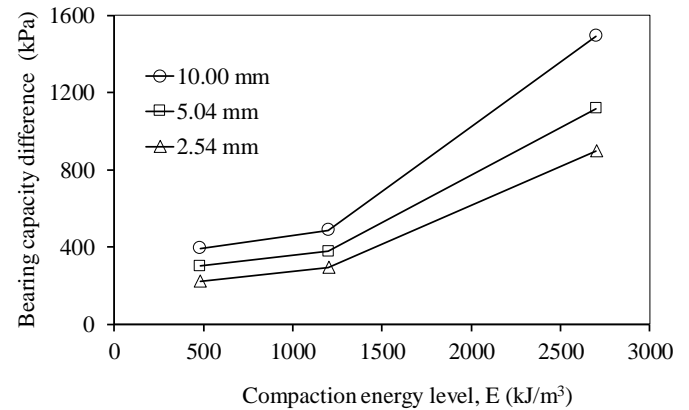


Fig. 5. The influence of compaction energy on bearing capacity improvement of clay reinforced by geotextile and 1.5cm thickness of sand cushion layer

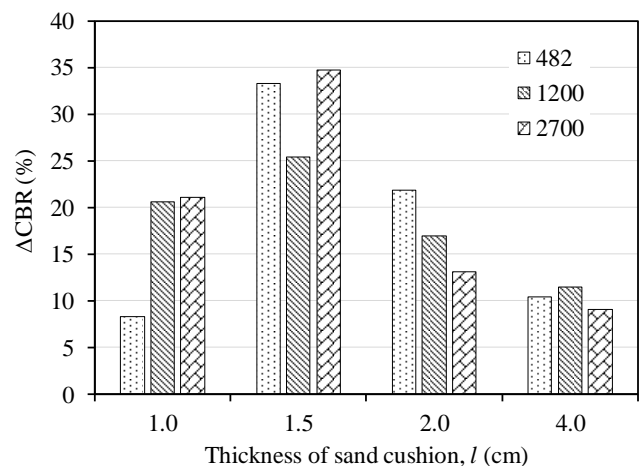


Fig. 6. The CBR improvement of clay specimens reinforced by geotextile with different thickness of sand cushion layer



## V. CONCLUSION

A series of CBR tests were performed to investigate the California Bearing Ratio, CBR of the soft clay specimens reinforced with nonwoven geotextile layers with sand cushion layer. The results demonstrate the benefit of a sand cushion layer covered by two geotextile layers that enhance the bearing capacity of the soft clay. The other conclusions are the following.

- The nonwoven geotextile with a sand cushion layer significantly improved the CBR value of clay. The reinforced specimens required sufficient deformation to mobilize the shear strength from soil-reinforcement interaction to improve the bearing capacity of the reinforced specimens. The more penetration of piston, the more bearing capacity enhancement was.
- To obtain the highest CBR and bearing capacity difference of the reinforced specimens, the optimum thickness of sand cushion was 1.5cm, of which the ratio h/B was about unity. In that case, the CBR value of the reinforced clay specimen could be raised up to 25.5-34.8% under the range of 486-2700 kJ/m<sup>3</sup> of compaction energy.
- The reinforced clay specimens should be compacted using high compaction energy. The denser the clay was, the more CBR and bearing capacity improvement could be achieved.

Last, the significant enhancement of the bearing capacity of the soft clay reinforced by two nonwoven geotextile layers with a sand cushion layer approved the potential application of the proposed method for pavement design and GRS structures using the reinforced clay with a thin sand cushion layer.

## ACKNOWLEDGMENT

The author sincerely appreciates the constructive comments and feedback by the anonymous reviewers.

## REFERENCES

- [1] A. Sridharan, B. R. S. Murthy and K. Revanasiddappa, "Technique for using fine-grained soil in reinforced earth.," *J. Geotech. Eng. (ASCE)*, vol. 117:8, no. 1174, p. 1174–1190, 1991.
- [2] M. R. Abdi and M. Arjomand, "Pullout tests conducted on clay reinforced with geogrid encapsulated in thin layers of sand.," *Geotext. Geomem.*, vol. 29, no. 6, p. 588–595, 2011.
- [3] M. R. Abdi and A. R. Zandieh, "Experimental and numerical analysis of large scale pull out tests conducted on clays reinforced with geogrids encapsulated with coarse material.," *Geotext. Geomem.*, vol. 42, no. 5, p. 494–504., 2014.
- [4] N. Unnikrishnan, K. R. Jagopal and N. R. Krishnaswamy, "Behaviour of reinforced clay under monotonic and cyclic loading," *Geotext. Geomem.*, vol. 20, no. 2, p. 117–133., 2002.
- [5] K. H. Yang, W. M. Yalaw, and M. D. Nguyen, "Behavior of Geotextile-Reinforced Clay with a Coarse Material Sandwich Technique under Undrained Triaxial Compression," *International Journal of Geomechanics*, vol. 16, no. 3, pp. GM.1943-5622.0000611, 2015.
- [6] D. V. Raisinghani and B. V. S. Viswanadham, "Evaluation of permeability characteristics of a geosynthetic-reinforced soil through laboratory tests.," *Geotext. Geomem.*, vol. 28, no. 6, p. 579–588., 2010.
- [7] H. I. Ling, & F. Tatsuoka, "Performance of Anisotropic Geosynthetic-Reinforced Cohesive Soil Mass," *Journal of Geotechnical Engineering*, vol. 120, no. 7, pp. 1166-1184, 1994.
- [8] R. M. Koerner, and D. Narejo, "Bearing Capacity of Hydrated Geosynthetic Clay Liners," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 121, no. 1, p. 82–85, 1995.
- [9] F. Moghaddas-Nejad and J. C. Small, "Effect of Geogrid Reinforcement in Model Track Tests on Pavements.," *Journal of Transportation Engineering*, vol. 122, no. 6, p. 468–474., 1996.
- [10] A. Choudhary, K. Gill, J. Jha, and S. K. Shukla, "Improvement in CBR of the expansive soil subgrades with a single reinforcement layer," in *Proceedings of Indian Geotechnical Conference.*, New Delhi, India, 2012.
- [11] N. Keerthi, and S. Kori, "Study on Improvement of Sub Grade Soil using Soil-Reinforcement Technique," *International Journal of Applied Engineering Research*, vol. 13, no. 7, pp. 126-134, 2018.
- [12] M. Singh, A. Trivedi, and S. Kumar Shukla, "Strength Enhancement of the Subgrade Soil of Unpaved Road with Geosynthetic Reinforcement Layers," *Transportation Geotechnics*, vol. 19, pp. 54-60, 2019.
- [13] M. A. Kamel, S. Chandra, and P. Kumar, "Behaviour of Subgrade Soil Reinforced with Geogrid," *International Journal of Pavement Engineering*, vol. 5, no. 4, pp. 201-209, 2004.
- [14] C. A. Adams, Y. A. Tuffour, and S. Kwofie, "Effects of Soil Properties and Geogrid Placement on CBR Enhancement of Lateritic Soil for Road Pavement Layers," *American Journal of Civil Engineering and Architecture*, 2016.
- [15] D. M. Carlos, M. Pinho-Lopes and M. L. Lopes, "Effect of Geosynthetic Reinforcement Inclusion on the Strength Parameters and Bearing Ratio of a Fine Soil," *Procedia Engineering*, vol. 143, p. 34–41, 2016.
- [16] M. R. Abdi, A. Sadrnejad, and M. A. Arjomand, "Strength enhancement of clay by encapsulating geogrids in thin layers of sand," *Geotext. Geomem.*, vol. 27, no. 6, p. 447–455., 2009.
- [17] D1883, "Standard Test Method for California Bearing Ratio (CBR) of Laboratory-Compacted Soils," in *ASTM*, West Conshohocken, PA, USA., ASTM International.
- [18] U. Rajesh, S. Sajja, and V. K. Chakravarthi, "Studies on Engineering Performance of Geogrid Reinforced Soft Subgrade," *Transportation Research Procedia*, vol. 17, p. 164–173., 2016.
- [19] M. D. Nguyen, K. H. Yang, S. H. Lee, C. S. Wu, and M. H. Tsai, "Behavior of nonwoven geotextile-reinforced sand and mobilization of reinforcement strain under triaxial compression," *Geosynthetics International*, vol. 20, no. 3, pp. 207-225, 2013.

# An Improvement of Maximum Power Point Tracking Algorithm Based on Particle Swarm Optimization Method for Photovoltaic System

Xuan Truong Luong

Department of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam  
xuantruongpdl@gmail.com

Van Hien Bui<sup>1,2</sup>

<sup>1</sup>Department of Electrical and Electronics Engineering  
Ho Chi Minh City University of Food Industry  
<sup>2</sup>Research student at HCMC University of Technology and Education  
Ho Chi Minh City, Viet Nam  
hienbv.ncs@hcmute.edu.vn or buivanhientb@gmail.com

Duc Tri Do

Department of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam  
tridd@hcmute.edu.vn

Thanh Hai Quach

Department of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam  
haiqt@hcmute.edu.vn

Viet Anh Truong

Department of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam  
anhvtv@hcmute.edu.vn

**Abstract**— Partial shading is the cause of the reduction of the output power of the photovoltaic (PV) system due to changes in its P-V characteristic curve. Global Maximum Power Point Tracking (GMPPT) in more complex and multiple peaks conditions is the biggest challenge for current MPPT techniques to improve the performance of the system. This article introduces an improved method based on the traditional Particle Swarm Optimization (I\_PSO) algorithm to increase the convergence speed in a constantly changing and complex environment. The study not only considers the influence of the best location of the individual and the swarm but also focuses on the experience of the neighboring individuals with a better position to avoid the local extreme trap. In addition to that, a boost converter uses to simulate the proposed algorithm applying PSIM software. The simulating results with those previously under the same operating conditions showed the superiority of the proposed approach in improving the efficiency of the photovoltaic system.

**Keywords**—partial shading, photovoltaic system, global maximum power point tracking, solar array.

## I. INTRODUCTION

In general, solar energy has been becoming a useful alternative fuel source for world energy security in recent years. The use of photovoltaic systems increases significantly because of its outstanding advantages such as low fuel and maintenance costs, environmentally friendly, and almost endless energy sources [1,2]. However, the expense of a PV plant and conversion equipment to reach the limit of the electricity system is a significant challenge in the development and use of them. Besides that, the PV characteristic curves depend on operating environment conditions as solar radiation and temperature, which can occur due to multiple reasons such as buildings, trees, or passing clouds, birds, and dust deposition...called partial shading. They are causes of energy losses in PV power generators. The traditional MPPT techniques such as P&O (Perturb and Observation), InC (Incremental Conductance) are significantly effective under uniform conditions but

inaccurate under partial shading conditions [3]. To overcome these drawbacks and improve the performance in MPPT control techniques. The improvement methods based on two algorithms also introduced recently as I\_InC (Improved Incremental Conductance) [4-6], VSSP&O (Variable Step Size Perturb and Observe) [7]. The optimization algorithms and their improvements based on the natural behavior of the swarm such as PSO (Particle Swarm Optimization), OD\_PSO (Differential Particle Swarm Optimization), LPSO (Leader Particle Swarm Optimization), EL\_PSO (Enhanced Leader Particle Swarm Optimization), MPSO (Multicore Particle Swarm Optimization) [8-12]. Another group is using hybrid methods that have emerged by combining two or more approaches in a solution to further enhance the performances as PSO-OCC (Particle Swarm Optimization Combined with one Cycle Control), SA-PSO (Simulated Annealing with Particle Swarm Optimization), INC-FFA (Firefly Algorithm with Incremental Conductance), PSO-P&O (Particle Swarm Optimization with Perturb and Observation), PSO-SFLA (Particle Swarm Optimization with Shuffled Frog Leaping Algorithm), and ABC-P&O (Artificial Bee Colony) [13-18]. These algorithms have achieved significant effects in solving the multi-peak GMPPT problem. However, the convergence speed, performance, applicability, and complexity are still barriers to the above solutions.

This paper proposes an improved method based on PSO to increase the convergence speed in GMPPT under continuously changing and complicated operation conditions. Compared to other MPPT techniques, the proposed algorithm has faster convergence speed, lower control cost, and greater efficiency by adding influence coefficients from neighboring particles with a better position. The remainder of this paper is as follows. Section II discusses the mathematical modeling of Solar PV and the effect of partial shading on the performance of the PV array. Section III proposed the I\_PSO algorithm and how to use it for improving the MPPT performance. In section IV, the simulating results show and compare with



other algorithms. Finally, Section 5 presents the conclusion of the work.

## II. PV SIMULATION MODEL AND THE EFFECT OF PARTIAL SHADING ON ITS CHARACTERISTICS

### A. PV Simulation Model.

The equivalent circuit for a photovoltaic cell introduced in Fig. 1 consists of a current source driven by sunlight in parallel with a real diode and both parallel and series resistances  $R_p$ ,  $R_s$  [19-22].

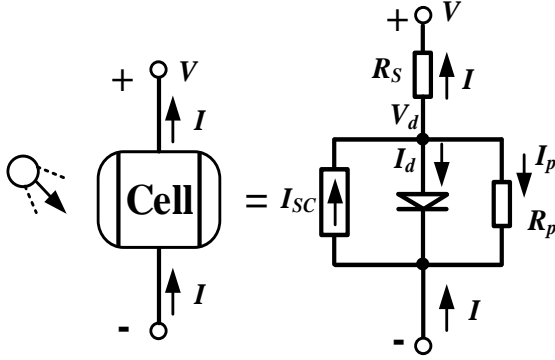


Fig. 1. PV cell equivalent circuit.

The current output of PV cell is:

$$I = I_{sc} - I_0 \left\{ e^{\frac{q(V+I.R_s)}{kT}} - 1 \right\} - \frac{V + I.R_s}{R_p} \quad (1)$$

Where:  $V$ ,  $I$  are output voltage (V) and current (A) of PV, respectively;  $I_{sc}$  is short circuit current (A);  $I_0$  is the reverse saturation current of diode (A);  $q = 1,602.10^{-19}$  (C);  $k = 1,381.10^{-23}$  (J/K);  $T$  (°K);  $R_s$ ,  $R_p$  (Ω).

### B. Effect of partial shading on the PV system.

The simulation and discussion in the article base on a series connection of 5 PV modules type PHM60W36 with its parameters are 37.5 (V), 33.3 (A), and 330 (W) at MPPT under operating conditions ( $1000\text{W/m}^2$  at  $25^\circ\text{C}$ ).

Previous studies have shown that, under uniform conditions, the P-V and I-V Characteristics Curves of the PV system are not different from those of a PV cell. Conversely, when the PV module operated under heterogeneous conditions like for partial shading cases caused by natural phenomena will increase multiple local maxima in the array's PV characteristics, which are capable of making it complex to determine GMPP and affect the system' output power (Fig. 2). Accordingly, the use of MPPT algorithms is necessary in this case [23].

The configuration of the proposed system simulates under partial shading conditions while only one case operates under a uniform environment. The data for changing solar radiation on the PV system shows in Table 1. When the irradiance is uniform, there is only an MPP (with maximum output power) exists on the P-V characteristic (Fig. 2). Meanwhile, the remaining cases have lower output power. It' characteristic curves also exhibit more LMPP. These show that the PV output power is affected by the level of shading on the modules. In other words, the radiation value received by modules will determine the whole system's output power. On

the other hand, accurately determine GMPP under multiple-peaks conditions is the biggest challenge for MPPT algorithms to improve the operating efficiency of the PV system. Because of under partial shading conditions as well as instantaneous environmental changes, it requires more efficient algorithms and higher convergence speed.

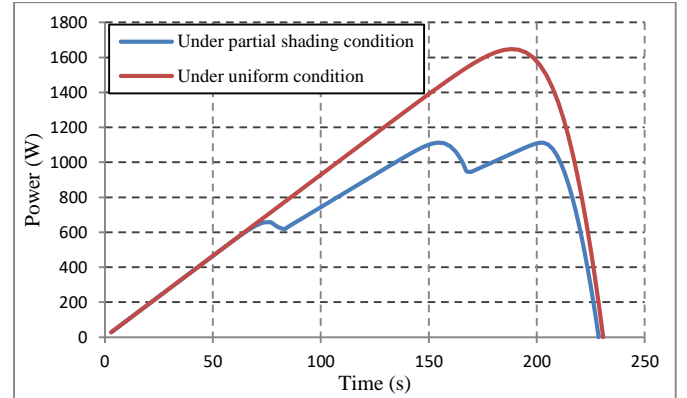


Fig. 2. P-V characteristic under different conditions.

TABLE I. THE PROPOSED CASES SIMULATE THE SYSTEM.

Case	Radiation Values Per Modules ( $\times 100 \text{ W/m}^2$ )				
	1	2	3	4	5
1	10	10	10	10	10
2	10	9	8	7	6
3	5	4	3	2	10
4	9	8	7	6	5
5	4	3	2	10	9
6	8	7	6	5	4
7	3	2	10	9	8
8	7	6	5	4	3
9	2	10	9	8	7
10	6	5	4	3	2

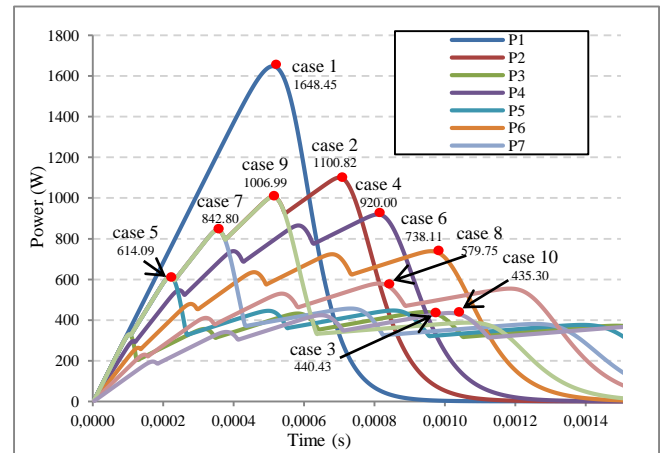


Fig. 3. GMPP under partial shading conditions

## III. THE PROPOSED SOLUTION

### A. DC/DC Converter.

A boost converter links between the PV array and the load to control the system working at the MPP is a DC/DC power converter. It is a class of switched-mode power supply containing at least two semiconductors (a diode and a transistor) and at least one energy storage element: a capacitor, inductor, or both in combination. It' diagram presents in Fig. 4, which is characterized by its duty

cycle  $D$  ( $0 \leq D \leq 1$ ) that gives the ratio between the input and the output voltage when the conduction is continuous [24].

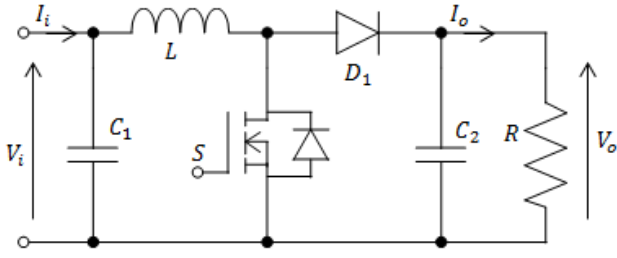


Fig. 4. Boost converter

The relationship between the input and output voltage depends on  $D$  is expressed by the following equation:

$$V_i = (1 - D)V_o \quad (2)$$

Where  $D$ ,  $V_i$ ,  $V_o$  are respectively the duty cycle, PV input voltage, and the output voltage of the Boost converter.

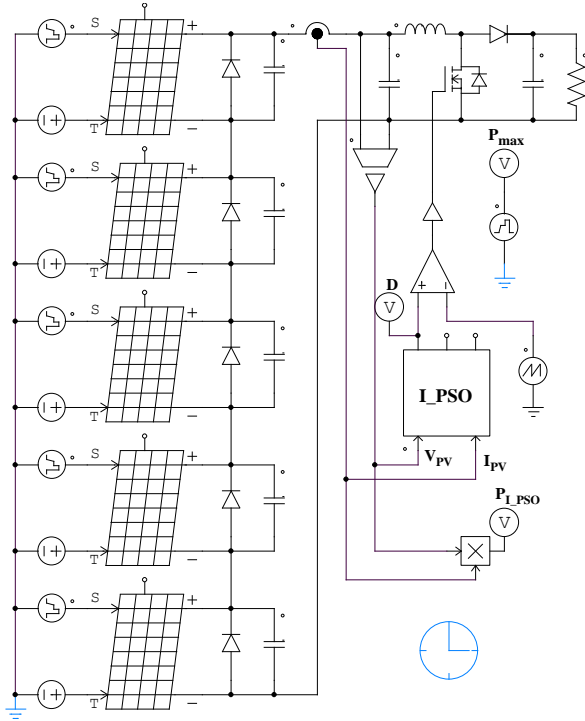


Fig. 5. Simulation model of the proposed system

With a load value is  $R = 120\Omega$ , maximum input voltage  $V_i = 200V$ , switching frequency  $f = 50kHz$ , and a duty cycle is between 20-80% to ensure that it does not exceed the limit of the semiconductor switch. The determining of inductor value according to the continuous and discontinuous conduction modes by the following equation:

$$L = \frac{DV_i}{2I_m f} \quad (3)$$

Where  $I_m$  is the maximum power current of the PV system.

The output capacitor value to decrease output voltage ripple is:

$$C = \frac{D}{Rfr} \quad (4)$$

Where: output voltage ripple  $r$  is 1%.

Therefore, the inductors and capacitors' value for boost converter used for the proposed method, and other parameters lists in Table 2. The system is simulated in the PSIM environment and has the configuration shown in Fig. 5

TABLE II. THE SPECIFICATIONS OF THE IMPLEMENTED BOOST CONVERTER

The parameters	Set value
Input voltage	$V_i = 80 - 200V$
Output voltage	$V_o = 400V$
Output current	$I_o = 5A$
Output power	$P_o = 2kW$
Switching frequency	$f = 50kHz$
Output voltage ripple	$r \leq 1\%$

### B. I\_PSO Implemented for MPPT.

PSO is a swarm intelligence optimization algorithm based on two main principles, i.e., to follow the best performing particle ( $G_{best}$ ), and to move towards the best conditions found by the particle itself ( $P_{best}$ ) [25, 26]. The I\_PSO introduced in this article is also an improved version, in which the particles can avoid LMPP traps to give better positioning in the search space and faster convergence speed. The proposed study not only considers the influence from the best location of the individual ( $P_{best}$ ), overall experience ( $G_{best}$ ), and the present movement of the particles but also focused on the effect of the neighboring individuals with a better position ( $P_e$ ), which used to decide their next values in the search space to avoid the local extreme trap.

Mathematically, the concepts of I\_PSO can be expressed as follows:

$$v_i^{k+1} = w_i v_i^k + c_1 r_1 (P_{best,i} - x_i^k) + c_2 r_2 (G_{best} - x_i^k) + c_3 r_3 (P_e - x_i^k) \quad (5)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (6)$$

Where  $x_i$ ,  $v_i$  are the position and velocity of  $i$  particle;  $k$  denotes the iteration number;  $w_i$  is the inertia weight;  $r_1$  and  $r_2$  are random variables uniformly distributed within  $[0, 1]$ ; and  $c_1$ ,  $c_2$  are the cognitive and social coefficient, respectively. The  $P_{best,i}$  variable uses to store the best position that a particle has found so far ( $i_{th}$ ), while the  $G_{best}$  is the best position of all the particles.

The significant difference in Eq. (5) is the influence coefficient  $c_3$  and a randomly-generated random number  $r_3$  between 0 and 1. Whereas,  $P_e$  represents the best position of an expert partial, which is better than  $P_{best}$  but not equal to  $G_{best}$ . Under this condition, when the  $G_{best}$  value of the swarm is determined, it is also its  $P_{best}$  local, which is likely reaching a velocity value equal to zero ( $v = 0$ ) when added in the update function. As a result, the survey range will reduce because this position does not change in the next cycle. The addition of third-place effects out of  $P_{best}$  and  $G_{best}$  parameters is necessary to ensure that it will survey all of the swarm.

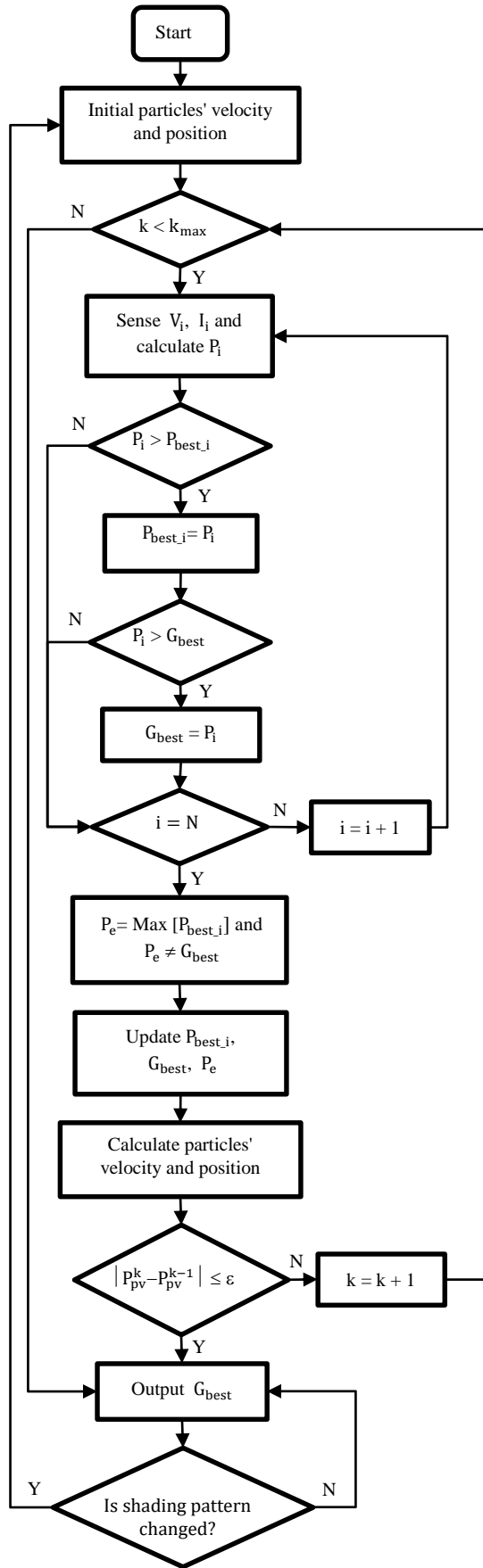


Fig. 6. Flowchart of the proposed algorithm

The complete flowchart of the proposed method illustrated in Fig. 6, the main blocks are described in detail as follows:

Step 1: In the proposed system, the selecting duty cycle  $D$  and delta  $D$  values of the DC-DC converter like the particle' position and velocity in the search space, respectively. Meanwhile, the fitness value evaluation function sets as the generated power of the whole system  $P_{PV}$ . The number of swarms in this paper is equal to six may be increased convergence time. However, the probability of overcome local extremes will significantly higher.

Step 2: Although particles can be placed on fixed positions or in the space randomly. In this paper, the particles' initialization value is permanent, which limits between 0.2 and 0.8 the search space with equal distances. It means that this is a region with the highest GMPP potential and also within the boost converter's operating range.

Step 3: After the digital controller output the PWM command according to the position of the  $i_{th}$  particle, which represents the duty cycle command, the voltage  $V_{PV}$ , and current  $I_{PV}$  are defined. These values are the basis of calculating the fitness evaluation of the  $i_{th}$  particle.

Step 4: If the fitness value of the  $i_{th}$  particle is better the previous the best value ( $P_{best,i}$ ), sets the current value as the new  $P_{best,i}$ . And then, this value can be utilized as a new  $G_{best}$  if it is greater than the  $G_{best}$  value, which is the best fitness value of all the particles in history (or not).

Step 5: Select  $P_e$  value as the expert particle, which is the best value of all  $P_{best,i}$  without  $G_{best}$  after each loop. Then, update all  $P_{best,i}$ ,  $G_{best}$ , and  $P_e$  values.

Step 6: Update velocity and position of all the particles in the swarm by using Eq. (5) and (6).

Step 7: When the algorithm reaches the maximum number of iterations, or as soon as the output power of the PV system is not significantly different between two consecutive loops, it will stop and output the  $G_{best}$  solution. Convergence criteria determine according to Eq. (7) then it is utilized to detect the insolation change and shading pattern changes.

$$|p_{pv}^k - p_{pv}^{k-1}| \leq \varepsilon \quad (7)$$

#### IV. SIMULATION RESULTS AND DISCUSSION

The parameter selection in the method to achieve the best simulating results plays an important role. According to the characteristics of the PV system uses in this study, the setting values of the proposed algorithm listed in Table 3, while Table 1 shows the simulation cases under partial shade conditions.

In the PSIM environment, the PV panel can be implemented using the physical model of the solar cell in the renewable energy package and structured in Fig. 5. The MPPT simulation results for the proposed system summarized in Fig. 7, which is known as the measured output power waveform of the PV system when treated by the proposed algorithm in the PSIM environment. Meanwhile, Fig. 8 shows the result of the generated output power waveforms when the irradiance changes. The MPPT simulation results using the proposed algorithm compared

with the maximum available power of the PV module under a continuous shading pattern, which is extracted from Fig. 3 to determine the tracking effectiveness of this algorithm. Besides that, Figs 9 and 10 show a comparison of the maximum power point tracking speed and performance of the proposed algorithm compared with traditional PSO. Meanwhile, Table 4 introduces the MPPT convergence speed and performance of all the cases presented above.

From Fig. 7, the proposed algorithm always extracts GMPPT under different shading conditions with convergence time between 0.22s and 0.39s in cases 9 and 10, respectively. Even though under simulation conditions changed instantaneously, the system is still capable of MPPT quickly with significant performance (Fig. 8). The aggregated results in Table 5 also show two vital issues: Firstly, the proposed algorithm can reach maximum efficiency of 100% in a few cases, but the average efficiency is always greater than 99%. Therefore, the proposed algorithm is as effective as the other optimal algorithms introduced recently. Secondly, the convergence speed of the algorithm is 0.22s when the measured MPPT tracking efficiency is 99.95% in the case of 9, which is a remarkable advantage compared to other modified and improved versions. The obtained simulation results in this paper are compared with other MPPT techniques under the same operating conditions and presented in Table 5. These positives show the superiority of the proposed method compared with previous techniques in both convergence speed and GMPPT tracking efficiency. Last but not least, although the convergence speed of the proposed algorithm significantly improved, the MPPT performance is not decreased, which listed in Figs 9 and 10, respectively.

TABLE III. THE I\_PSO ALGORITHM PARAMETERS

The parameters	Set value
Population sizes (N)	6
Maximum iteration ( $k_{max}$ )	100
Acceleration factors ( $c_1; c_2$ )	0.1; 0.5
Influence coefficient ( $c_3$ )	0.35
Inertia weight ( $w_i$ )	0.07
Random variables ( $r_1, r_2, r_3$ )	[0 1]
Sampling time	0.2 (s)

TABLE IV. THE MPPT PERFORMANCE AND CONVERGENCE SPEED

Case	P <sub>max</sub> (W)	P <sub>i_pso</sub> (W)	$\eta$ (%)	T (s)
1	1648.45	1648.40	100.00	0.22
2	1100.82	1100.73	99.99	0.30
3	440.43	440.39	99.99	0.31
4	920.00	917.93	99.78	0.31
5	614.09	613.39	99.89	0.29
6	738.11	737.75	99.95	0.31
7	842.80	842.70	99.99	0.28
8	579.75	579.73	100.00	0.38
9	1006.99	1006.51	99.95	0.22
10	435.30	434.89	99.91	0.39

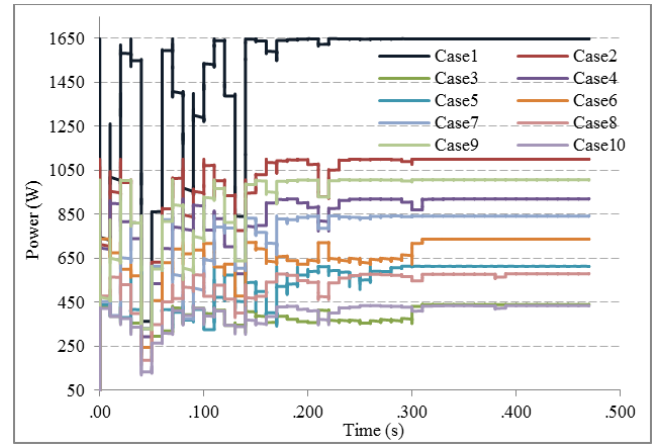


Fig. 7. Measured PV power waveforms under MPPT process.

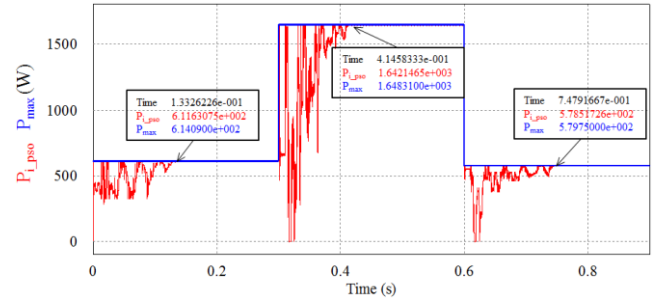
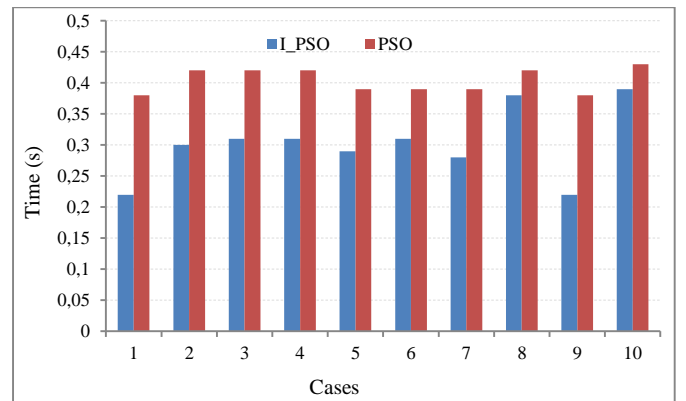


Fig 8. PV power waveforms when the irradiance changes

TABLE 5. COMPARISON OF GMPPT TECHNIQUES

MPPT Technique	$\eta$ (%)	T (s)	MPPT Technique	$\eta$ (%)	T (s)
I_PSO	100	0.22	L_PSO [10]	99.99	0.35
OD_PSO [9]	97.74	1.86	INC_FFA [15]	99.99	0.38
MPSO [12]	98.92	1.3	SA_PSO [14]	-	0.13
PSO [10]	99.83	0.85	PSO_P&O [16]	-	0.9
PSO_SFLA [17]	-	3.15	P&O [10]	99.95	0.52
ABC_P&O [18]	99.93	0.08	PSO_OCC [13]	100	1.2

Fig 9. Comparison of convergence speed with and without  $c_3$  coefficients

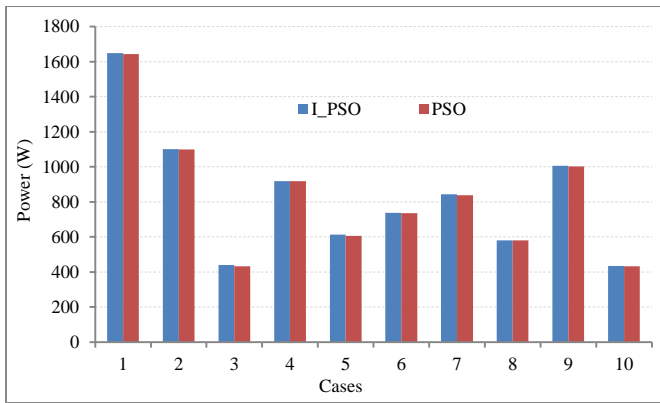


Fig 10. Comparison of MPPT performance with and without  $c_3$  coefficients

## V. CONCLUSION

I\_PSO is an improved version based on the traditional PSO algorithm introduced in this article, which not only has a fast convergence speed but also has an outstanding MPPT tracking efficiency under different operating conditions. The simulation results show that the performance of the proposed algorithm is likely reaching 100% at 0.22s, which are higher than previous improvements to the traditional PSO algorithm. It is due to the addition of an influencing factor in the velocity equation of the classical PSO algorithm to improve the operating efficiency of the PV system. Conclusion, I\_PSO is a simple and efficient technique, which is capable of escaping from LMPP traps and extracting the optimization power at high speed under partial shading condition as well as instantaneous environmental changes.

## REFERENCES

- [1] Ehsanul, Pawan Kumar, Sandeep, Adedeji A. Adelodun, Ki-Hyun Kim, Solar energy: Potential and future prospects, Renewable and Sustainable Energy Reviews 82 (2018) 894–900.
- [2] Al-Saidi, M., & Lahham, N. Solar energy farming as a development innovation for vulnerable water basins. Development in Practice, (2019).
- [3] Saleh Elkelani Babaa, Matthew Armstrong, Volker Pickert: Overview of Maximum Power Point Tracking Control Methods for PV Systems, Journal of Power and Energy Engineering, 2014, 2, 59–72.
- [4] Duy C. Huynh: An Improved Incremental Conductance Maximum Power Point Tracking Algorithm for Solar Photovoltaic Panels, International Journal of Science and Research, Volume 3 Issue 10, October 2014.
- [5] Nazih Moubayed, Ali El-Ali, Rachid Outbib: A comparison of two MPPT techniques for PV system, WSEAS Transactions on Environment and Development, Issue 12, Volume 5, December 2009.
- [6] Tawfik Radjai, Lazhar Rahmani, Saad Mekhilef, Jean Paul Gaubert; Implementation of a modified incremental conductance MPPT algorithm with direct control based on a fuzzy duty cycle change estimator using dSPACE, Solar Energy 110 (2014) 325–337
- [7] Riby John, Sheik Mohammed S, Richu Zachariah, Variable Step Size Perturb and Observe MPPT Algorithm for Standalone Solar Photovoltaic System, IEEE International Conference On Intelligent Techniques In Control, Optimization And Signal Processing, 2017
- [8] Chun-Liang Liu, Yi-Feng Luo, Jia-Wei Huang, Yi-Hua Liu: A PSO-based MPPT Algorithm for Photovoltaic Systems Subject to Inhomogeneous Insolation, SCIS-ISIS 2012, Kobe, Japan, November 20–24, 2012
- [9] Li, H., Yang, D., Su, W., Lü, J., Yu, X., 2018b. An overall distribution particle swarm optimization MPPT algorithm for photovoltaic system under partial shading. IEEE Trans. Ind. Electron. 66 (1), 265–275
- [10] Ram JP, Rajasekar N. A new robust, mutated and fast tracking LPSO method for solar PV maximum power point tracking under partial shaded conditions. Appl Energy 2017;201:45–59
- [11] Pratik Shantaram Gavhane, Smriti Krishnamurthy, Ridhima Dixit, J. Prasanth Ram, N. Rajasekar; EL-PSO based MPPT for Solar PV under Partial Shaded Condition, Energy Procedia 117 (2017) 1047–1053.
- [12] Chao RM, Nasirudin A, Wang IK, Chen PL. Multicore PSO operation for maximum power point tracking of a distributed photovoltaic system under partially shading condition. Int J of Photoenergy (2016) 1–19.
- [13] Anoop, K., Nandakumar, M., A novel maximum power point tracking method based on particle swarm optimization combined with one cycle control. Proceedings of the International Conference on Power, Instrumentation, Control and Computing (PICCC). Thrissur 2018.
- [14] Guan T, Zhuo F. An improved SA-PSO global maximum power point tracking method of photovoltaic system under partial shading conditions. In: Proceedings of the IEEE conference on environment and electrical engineering. Italy; 2017
- [15] Shi JY, Ling LT, Xue F, Qin ZJ, Li YJ, Lai ZX, et al. Combining incremental conductance and firefly algorithm for tracking the global MPP of PV arrays. J Renew Sustain Energy 2017;9(2):1–19.
- [16] Hanafiah S, Ayad A, Hehn A, Kennel R. A hybrid MPPT for quasi-Z-source inverters in PV applications under partial shading condition. In: Proceedings of the 11th IEEE international conference on compatibility, power electronics and power engineering; 4–6 April 2017.
- [17] Nie X, Nie H. MPPT control strategy of PV based on improved shuffled frog leaping algorithm under complex environments. J Control Sci Eng 2017;2017:1–12
- [18] Pilakkat, D., Kanthalakshmi, S., An improved P&O algorithm integrated with artificial bee colony for photovoltaic systems under partial shading conditions. Sol. Energy 178, 37–47, 2019.
- [19] Neeraj Priyadarshi, Sanjeevikumar Padmanaban, Lucian Mihet Popa, Frede Blaabjerg and Farooque Azam, Maximum Power Point Tracking for Brushless DC Motor-Driven Photovoltaic Pumping Systems Using a Hybrid ANFIS-FLOWER Pollination Optimization Algorithm, (2018), 11(5), 1–16
- [20] El-Helw HM, Magdy A, Marei MI. A hybrid maximum power point tracking technique for partially shaded photovoltaic arrays. IEEE Access 2017;5:11900–8.
- [21] Chakkarapani M, Raman GP, Raman GR, Ganesan SI, Chilakapati N. Fireworks enriched P&O algorithm for GMPPT and detection of partial shading in PV systems. IEEE Trans Power Electron 2017;32(6):4432–43.
- [22] Mohanty S, Subudhi B, Ray PK. A Grey Wolf assisted perturb & observe MPPT algorithm for a PV system. IEEE Trans Energy Convers 2016;32(1):340–7.
- [23] Faiza Belhachat, Cherif Larbes, A review of global maximum power point tracking techniques of photovoltaic system under partial shading conditions, Renewable and Sustainable Energy Reviews 92 (2018) 513–553.
- [24] Preti Tyagi, V.C. Kotak, V. P. Sunder Singh, Design High Gain DC-DC Boost Converter with Coupling inductor and Simulation in PSIM, International Journal of Research in Engineering and Technology, Volume: 03 Issue: 04, Apr-2014
- [25] Vapnik, V.N. 1995. The Nature of Statistical Learning Theory. 2nd ed.; Springer-Verlag, New York.
- [26] Kennedy, J., Eberhart, R.C., Shi, Y. 2001. Swarm Intelligence, Morgan Kaufmann, San Francisco, USA.



# Static Analysis of Sandwich Plates using ES-MITC3 Elements based on the Third-order Shear Deformation Layerwise Theory

Thanh Chau-Dinh

Faculty of Civil Engineering  
HCMC University of Technology and  
Education  
01 Vo Van Ngan Street, Thu Duc  
District, Ho Chi Minh City, Vietnam  
chdthanh@hcmute.edu.vn

Loi Dang-Huu

Kim Hung Thinh Consultant Design  
Construction Co., Ltd  
Vinh Tan Commune, Tan Uyen Town,  
Binh Duong Province, Vietnam  
danghuuloispkt2010@gmail.com

Jin-Gyun Kim

Department of Mechanical Engineering  
Kyung Hee University  
Deogyong-daero, Giheung-gu,  
Yongin-si, Gyeonggi-do, 17104,  
Republic of Korea  
jngyun.kim@khu.ac.kr

**Abstract**—In this paper, a 3-node triangular plate element is developed for static analysis of sandwich plates. The  $C^0$ -plate element attenuates the shear-locking phenomenon by approximating the transverse shear strains according to the mixed interpolation of tensorial components technique (MITC3). The constant strain fields within each MITC3 plate element are improved by using the edge-based smoothed approach (ES), in which the strain fields are averaged on domains of two adjacent elements. Based on the layerwise theory with the displacement fields for each layer described by the third-order shear deformation theory (TSDT), the formulation of the ES-MITC3 plate element is derived for the static analysis of sandwich plates. The accuracy and efficiency of the proposed approach are verified through some benchmark sandwich plates subjected to sinusoidal or uniform distributed loads.

**Keywords**—ES-MITC3 plate element, layerwise theory, TSDT, sandwich plates

## I. INTRODUCTION

Sandwich plates are special cases of multilayered composite plates, which consist of two thin but stiff skins adhesively bonding with a lightweight but thick core. The sandwich plates are widely used in many engineering structures because of their advantageous properties of high bending strength-to-weight ratio, acoustical and thermal insulation [1]. Consequently, many researchers are interested in studying deformation theories and solving methods of the sandwich plates. Deformations of the sandwich plates can be predicted by the equivalent single-layer (ESL) theory or layerwise (LW) theory besides the 3-dimensional elastic theory which models each layer as a 3-dimensional deformable body [1]. The ESL theory replaces constitutive behaviors of all layers by an equivalent one of a single plate. In contrast, the LW theory considers deformations of each layer as separate plates but constrains the continuity of the deformations at the interfaces between layers. Among these theories, the LW theory can represent the deformations through the plate's thickness better than those given by the ESL theory and consumes less computational cost in comparison with the 3-dimensional elastic theory. Instead of using the first-order shear deformation theory (FSDT), the displacement fields in each layer in the LW theory can be described by the higher-order shear deformation theory (HSDT). By including such higher-order functions as third-order polynomial, trigonometric, hyperbolic, or exponential functions into the displacement fields, the HSDT improves the

transverse deformation and does not require shear correction factors as in the FSDT [2]. For simplicity, the LW theory in this study uses the third-order shear deformation theory (TSDT) for each layer, namely the TSD-LW theory, to predict the behaviors of the sandwich plates.

Although some numerical methods like the meshless [3] or isogeometric (IGA) [4] methods have been suggested, the finite element methods (FEM) are still popular and play an important role in solving engineering problems. For sandwich plates, many finite element formulations have been established based on the LW theory [5–7]. One of the simplest approaches is the development of  $C^0$ -type 3-node triangular plate elements. However, the pure  $C^0$ -type elements cannot eliminate the transverse shear strains when used to predict thin plates and cause the shear-locking phenomenon. To remedy this phenomenon, the transverse shear strains are separately interpolated according to the Discrete Shear Gap (DSG3) [8] or Mixed Interpolation Tensorial Components (MITC3) techniques [9]. Recently, Liu and Nguyen-Thoi [10] have proposed the smoothed FEM (SFEM) to reduce differences in the strain fields given by 3-node triangular plate elements. According to the SFEM, the strains fields are usually modified within elements, or domains of elements sharing common edges or nodes, respectively called the cell-based (CS-), edge-based (ES-) or node-based (NS-) FEM. The combination of the SFEM with the DSG3 plate elements to construct ES-DSG3 and CS-DSG3 elements [11, 12] have been developed for analysis of laminated composite and sandwich plates based on the FSD-LW and TSD-LW theories. Similarly, the ES-MITC3 element has also been proposed for static analysis of multilayered composite plates using the FSD-LW theory [13]. Therefore, in this paper, the ES-MITC3 element is formulated for 3-layer sandwich plates employing the TSD-LW theory.

In the next section, the displacement fields of the TSD-LW theory for 3-layer sandwich plates are briefly presented and then the ES-MITC3 element is constructed. Some sandwich plates are statically analyzed to verify the presented element in Section 3. In the last section, several conclusions are withdrawn.

## II. ES-MITC3 ELEMENT FOR SANDWICH PLATES USING TSDT FOR LAYERWISE THEORY

Consider a  $t$ -thickness sandwich plate, including 3 layers with the corresponding thicknesses  $t_1$ ,  $t_2$ , and  $t_3$ , subjected to transverse distributed load  $q$  as shown in Fig. 1a. Employing the TSDT to describe the displacement fields for each layer



and constraining displacement continuity at the interfaces between layers, the displacement fields of the 3-layer sandwich plate according to the TSD-LW theory are [12]

$$\begin{aligned} u^{(1)} &= -\left(\frac{t_2}{2} - \frac{t_2}{6}\right)\beta_x^{(2)} + \frac{t_2}{6}\phi_x^{(2)} - \frac{t_1}{3}\beta_x^{(1)} + \frac{t_1}{6}\phi_x^{(1)} + (z_1 - c_1 z_1^3)\beta_x^{(1)} - c_1 z_1^3 \phi_x^{(1)} \\ v^{(1)} &= -\left(\frac{t_2}{2} - \frac{t_2}{6}\right)\beta_y^{(2)} + \frac{t_2}{6}\phi_y^{(2)} - \frac{t_1}{3}\beta_y^{(1)} + \frac{t_1}{6}\phi_y^{(1)} + (z_1 - c_1 z_1^3)\beta_y^{(1)} - c_1 z_1^3 \phi_y^{(1)} \\ u^{(2)} &= (z_2 - c_2 z_2^3)\beta_x^{(2)} - c_2 z_2^3 \phi_x^{(2)}; v^{(2)} = (z_2 - c_2 z_2^3)\beta_y^{(2)} - c_2 z_2^3 \phi_y^{(2)} \\ u^{(3)} &= \left(\frac{t_2}{2} - \frac{t_2}{6}\right)\beta_x^{(2)} - \frac{t_2}{6}\phi_x^{(2)} + \frac{t_3}{3}\beta_x^{(3)} - \frac{t_3}{6}\phi_x^{(3)} + (z_3 - c_3 z_3^3)\beta_x^{(3)} - c_3 z_3^3 \phi_x^{(3)} \\ v^{(3)} &= \left(\frac{t_2}{2} - \frac{t_2}{6}\right)\beta_y^{(2)} - \frac{t_2}{6}\phi_y^{(2)} + \frac{t_3}{3}\beta_y^{(3)} - \frac{t_3}{6}\phi_y^{(3)} + (z_3 - c_3 z_3^3)\beta_y^{(3)} - c_3 z_3^3 \phi_y^{(3)} \\ w^{(1)} &= w^{(2)} = w^{(3)} = w_0 \end{aligned} \quad (1)$$

in which,  $w_0$  is the deflection of the plate;  $u^{<k>}$ ,  $v^{<k>}$  and  $w^{<k>}$  are respectively the  $x$ -,  $y$ -,  $z$ -directional displacements;  $\beta_x^{<k>}$  and  $\beta_y^{<k>}$  are the rotations of the vector normal to the midplane;  $\phi_x^{<k>}$  and  $\phi_y^{<k>}$  are the wrapping variables;  $c_k = 4/(3t_k^2)$ ; and  $k = 1, 2, 3$ .

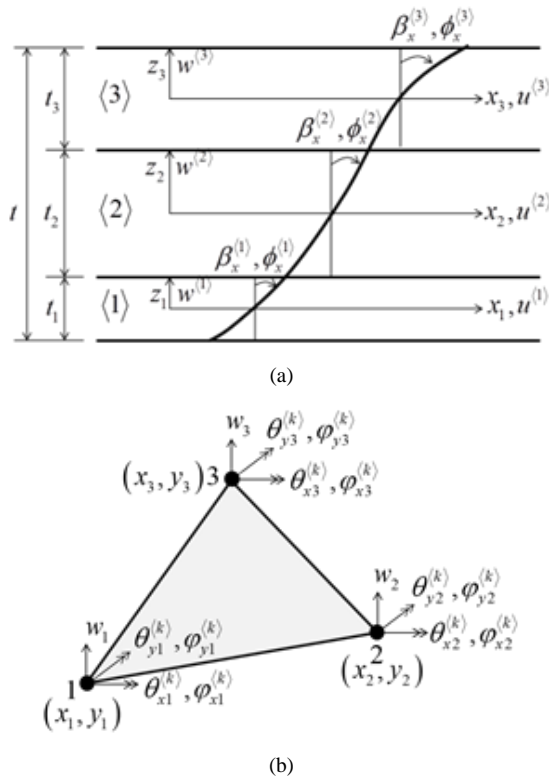


Fig. 1. (a) Displacement fields of a sandwich plate under transverse load; (b) Nodal displacements of 3-node triangular element for TSD-LW theory

The TSD-LW theory-based displacements of the sandwich plate are approximated by nodal displacements of 3-node triangular as follows [12]

$$\begin{aligned} w_0 &= \sum_{I=1}^3 N_I w_{0I}; \beta_x^{(k)} = \sum_{I=1}^3 N_I \theta_{xI}^{(k)}; \beta_y^{(k)} = -\sum_{I=1}^3 N_I \theta_{yI}^{(k)}; \\ \phi_x^{(k)} &= \sum_{I=1}^3 N_I \phi_{xI}^{(k)}; \phi_y^{(k)} = -\sum_{I=1}^3 N_I \phi_{yI}^{(k)} \end{aligned} \quad (2)$$

herein,  $w_{0I}$ ,  $\theta_{xI}^{<k>}$ ,  $\theta_{yI}^{<k>}$ ,  $\phi_{xI}^{<k>}$  and  $\phi_{yI}^{<k>}$  are the deflections, rotations and wrapping variables of node  $I$  as defined in Fig. 1b; and  $N_I$  are the  $C^0$  shape functions.

From the displacement fields in Eq. (1) and their approximations in Eq. (2), the strain fields in each layer can be expressed by the nodal displacements

$$\begin{bmatrix} \epsilon_{xx}^{(k)} & \epsilon_{yy}^{(k)} & \gamma_{xy}^{(k)} \end{bmatrix}^T = \mathbf{\epsilon}^{(k)} = \sum_{I=1}^3 \left( z_k \mathbf{B}_{1I}^{(k)} + \mathbf{B}_{2I}^{(k)} + \mathbf{B}_{3I}^{(k)} + z_k^3 \mathbf{B}_{4I}^{(k)} \right) \mathbf{d}_I \quad (3)$$

$$\begin{bmatrix} \gamma_{xz}^{(k)} & \gamma_{yz}^{(k)} \end{bmatrix}^T = \mathbf{\gamma}^{(k)} = \sum_{I=1}^3 \left( \mathbf{S}_{1I}^{(k)} + z_k^2 \mathbf{S}_{2I}^{(k)} \right) \mathbf{d}_I \quad (4)$$

with  $\mathbf{d}_I = [w_{0I} \ \theta_{xI}^{(1)} \ \theta_{yI}^{(1)} \ \phi_{xI}^{(1)} \ \phi_{yI}^{(1)} \ \theta_{xI}^{(2)} \ \theta_{yI}^{(2)} \ \phi_{xI}^{(2)} \ \phi_{yI}^{(2)} \ \theta_{xI}^{(3)} \ \theta_{yI}^{(3)} \ \phi_{xI}^{(3)} \ \phi_{yI}^{(3)}]^T$  and

$$\begin{aligned} \mathbf{B}_{1I}^{(1)} &= \begin{bmatrix} 0 & 0 & N_{I,x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}; \mathbf{B}_{1I}^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{1I}^{(3)} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (5)$$

$$\begin{aligned} \mathbf{B}_{2I}^{(1)} &= \begin{bmatrix} 0 & 0 & -\frac{t_1}{3} N_{I,x} & 0 & \frac{t_1}{6} N_{I,x} & 0 & -\frac{t_2}{2} N_{I,x} & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{2I}^{(2)} &= \begin{bmatrix} 0 & \frac{t_1}{3} N_{I,y} & 0 & -\frac{t_1}{6} N_{I,y} & 0 & \frac{t_2}{2} N_{I,y} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{2I}^{(3)} &= \begin{bmatrix} 0 & \frac{t_1}{3} N_{I,x} & -\frac{t_1}{3} N_{I,y} & -\frac{t_1}{6} N_{I,x} & \frac{t_1}{6} N_{I,y} & \frac{t_2}{2} N_{I,x} & -\frac{t_2}{2} N_{I,y} & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (6)$$

$$\begin{aligned} \mathbf{B}_{3I}^{(1)} &= \begin{bmatrix} 0 & 0 & \frac{t_2}{2} N_{I,x} & 0 & 0 & \frac{t_3}{3} N_{I,x} & 0 & -\frac{t_3}{6} N_{I,x} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{3I}^{(2)} &= \begin{bmatrix} 0 & 0 & -\frac{t_2}{2} N_{I,y} & 0 & 0 & -\frac{t_3}{3} N_{I,y} & 0 & \frac{t_3}{6} N_{I,y} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{3I}^{(3)} &= \begin{bmatrix} 0 & 0 & -\frac{t_2}{2} N_{I,x} & \frac{t_2}{2} N_{I,y} & 0 & -\frac{t_3}{3} N_{I,x} & \frac{t_3}{3} N_{I,y} & \frac{t_3}{6} N_{I,x} & -\frac{t_3}{6} N_{I,y} & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (7)$$

$$\begin{aligned} \mathbf{B}_{4I}^{(1)} &= -c_1 \begin{bmatrix} 0 & 0 & N_{I,x} & 0 & N_{I,x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{4I}^{(2)} &= -c_2 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{B}_{4I}^{(3)} &= -c_3 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (8)$$

$$\begin{aligned} \mathbf{S}_{1I}^{(1)} &= \begin{bmatrix} N_{I,x} & 0 & N_I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}; \mathbf{S}_{1I}^{(2)} = \begin{bmatrix} N_{I,x} & 0 & 0 & 0 & N_I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{1I}^{(3)} &= \begin{bmatrix} N_{I,x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (9)$$

$$\begin{aligned} \mathbf{S}_{2I}^{(1)} &= -3c_1 \begin{bmatrix} 0 & 0 & N_I & 0 & N_I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{2I}^{(2)} &= -3c_2 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{2I}^{(3)} &= -3c_3 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (10)$$

For a 3-node triangular element with the nodal coordinates  $(x_1, y_1)$ ,  $(x_2, y_2)$  and  $(x_3, y_3)$ , derivatives of the shape functions  $N_i$  with respect to  $x$  and  $y$  are determined by

$$N_{1,x} = \frac{b-c}{2A_e}; N_{1,y} = \frac{d-a}{2A_e}; N_{2,x} = \frac{c}{2A_e}; N_{2,y} = \frac{-d}{2A_e}; N_{3,x} = \frac{-b}{2A_e}; N_{3,y} = \frac{a}{2A_e} \quad (11)$$

with  $a = x_2 - x_1$ ,  $b = y_2 - y_1$ ,  $c = y_3 - y_1$ ,  $d = x_3 - x_1$  and  $A_e$ : area of the element.

The transverse shear strains purely derived from the displacement approximation in Eq. (2) will cause the shear-locking phenomenon when the sandwich plate's thickness reduces. To use the 3-node triangular element for analysis of both thin and thick plates, the transverse shear strains in Eq. (4) are separately interpolated and connected with the displacement approximation through tying points according to the MITC3 technique [9]. By employing one quadrature point, the transverse shear strains are explicitly related to the nodal displacements [14] as follows

$$\left[ \gamma_{xz}^{MITC3(k)} \quad \gamma_{yz}^{MITC3(k)} \right]^T = \gamma^{MITC3(k)} = \sum_{I=1}^3 \left( \mathbf{S}_{1I}^{MITC3(k)} + z_k^2 \mathbf{S}_{2I}^{MITC3(k)} \right) \mathbf{d}_I \quad (12)$$

in which,

$$\begin{aligned} \mathbf{S}_{11}^{MITC3(1)} &= \frac{1}{2A_e} \begin{bmatrix} b-c & (b-c)(b+c)/6 & A_e+(d-a)(b+c)/6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ d-a & -A_e-(b-c)(a+d)/6 & (a-d)(a+d)/6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{12}^{MITC3(1)} &= \frac{1}{2A_e} \begin{bmatrix} c & -bc/2+c(b+c)/6 & ac/2-d(b+c)/6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -d & -bd/2-c(a+d)/6 & -ad/2+d(a+d)/6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{13}^{MITC3(1)} &= \frac{1}{2A_e} \begin{bmatrix} -b & bc/2-b(b+c)/6 & -bd/2+a(b+c)/6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a & -ac/2+b(a+d)/6 & ad/2-a(a+d)/6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (13)$$

$$\begin{aligned} \mathbf{S}_{11}^{MITC3(2)} &= \frac{1}{2A_e} \begin{bmatrix} b-c & 0 & 0 & 0 & (b-c)(b+c)/6 & A_e+(d-a)(b+c)/6 & 0 & 0 & 0 & 0 \\ d-a & 0 & 0 & 0 & -A_e-(b-c)(a+d)/6 & (a-d)(a+d)/6 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{12}^{MITC3(2)} &= \frac{1}{2A_e} \begin{bmatrix} c & 0 & 0 & 0 & -bc/2+c(b+c)/6 & ac/2-d(b+c)/6 & 0 & 0 & 0 & 0 \\ -d & 0 & 0 & 0 & -bd/2-c(a+d)/6 & -ad/2+d(a+d)/6 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{13}^{MITC3(2)} &= \frac{1}{2A_e} \begin{bmatrix} -b & 0 & 0 & 0 & bc/2-b(b+c)/6 & -bd/2+a(b+c)/6 & 0 & 0 & 0 & 0 \\ a & 0 & 0 & 0 & -ac/2+b(a+d)/6 & ad/2-a(a+d)/6 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (14)$$

$$\begin{aligned} \mathbf{S}_{11}^{MITC3(3)} &= \frac{1}{2A_e} \begin{bmatrix} b-c & 0 & 0 & 0 & 0 & 0 & (b-c)(b+c)/6 & A_e+(d-a)(b+c)/6 & 0 & 0 \\ d-a & 0 & 0 & 0 & 0 & 0 & -A_e-(b-c)(a+d)/6 & (a-d)(a+d)/6 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{12}^{MITC3(3)} &= \frac{1}{2A_e} \begin{bmatrix} c & 0 & 0 & 0 & 0 & 0 & -bc/2+c(b+c)/6 & ac/2-d(b+c)/6 & 0 & 0 \\ -d & 0 & 0 & 0 & 0 & 0 & -bd/2-c(a+d)/6 & -ad/2+d(a+d)/6 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{13}^{MITC3(3)} &= \frac{1}{2A_e} \begin{bmatrix} -b & 0 & 0 & 0 & 0 & 0 & bc/2-b(b+c)/6 & -bd/2+a(b+c)/6 & 0 & 0 \\ a & 0 & 0 & 0 & 0 & 0 & -ac/2+b(a+d)/6 & ad/2-a(a+d)/6 & 0 & 0 \end{bmatrix} \end{aligned} \quad (15)$$

$$\begin{aligned} \mathbf{S}_{2I}^{(1)} &= -3c_1 \begin{bmatrix} 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1/3 & 0 & -1/3 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{2I}^{(2)} &= -3c_2 \begin{bmatrix} 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1/3 & 0 & -1/3 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \mathbf{S}_{2I}^{(3)} &= -3c_3 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1/3 & 0 & -1/3 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (16)$$

The strain fields of the MITC3 triangular element are constant on each one. Differences in the strains between elements can be reduced by applying the edge-based smoothed (ES) finite element method [10]. In this method, all strain fields of the MITC3 element are averaged on domains bounded by straight lines connecting 2 nodes of an edge with 2 centroids of 2 elements sharing the edge as illustrated in Fig. 2a. As a result, the strain fields of the ES-MITC3 element are

$$\begin{aligned} \tilde{\boldsymbol{\varepsilon}}^{(k)} &= \frac{1}{A_{ES}} \int_{A_{ES}} \boldsymbol{\varepsilon}^{(k)} dA = \sum_{I=1}^{N_{ES}} \left( z_k \tilde{\mathbf{B}}_{1I}^{(k)} + \tilde{\mathbf{B}}_{2I}^{(k)} + \tilde{\mathbf{B}}_{3I}^{(k)} + z_k^3 \tilde{\mathbf{B}}_{4I}^{(k)} \right) \mathbf{d}_I \\ \tilde{\boldsymbol{\gamma}}^{MITC3(k)} &= \frac{1}{A_{ES}} \int_{A_{ES}} \boldsymbol{\gamma}^{MITC3(k)} dA = \sum_{I=1}^{N_{ES}} \left( \tilde{\mathbf{S}}_{1I}^{MITC3(k)} + z_k^2 \tilde{\mathbf{S}}_{2I}^{MITC3(k)} \right) \mathbf{d}_I \end{aligned} \quad (17)$$

wherein,

$$\begin{aligned} \tilde{\mathbf{B}}_{1I}^{(k)} &= \frac{1}{A_{ES}} \sum_{e=1}^{E_{ES}} \frac{A_e}{3} \mathbf{B}_{1I,e}^{(k)}; \tilde{\mathbf{B}}_{2I}^{(k)} = \frac{1}{A_{ES}} \sum_{e=1}^{E_{ES}} \frac{A_e}{3} \mathbf{B}_{2I,e}^{(k)}; \\ \tilde{\mathbf{B}}_{3I}^{(k)} &= \frac{1}{A_{ES}} \sum_{e=1}^{E_{ES}} \frac{A_e}{3} \mathbf{B}_{3I,e}^{(k)}; \tilde{\mathbf{B}}_{4I}^{(k)} = \frac{1}{A_{ES}} \sum_{e=1}^{E_{ES}} \frac{A_e}{3} \mathbf{B}_{4I,e}^{(k)} \\ \tilde{\mathbf{S}}_{1I}^{MITC3(k)} &= \frac{1}{A_{ES}} \sum_{e=1}^{E_{ES}} \frac{A_e}{3} \mathbf{S}_{1I,e}^{MITC3(k)}; \tilde{\mathbf{S}}_{2I}^{MITC3(k)} = \frac{1}{A_{ES}} \sum_{e=1}^{E_{ES}} \frac{A_e}{3} \mathbf{S}_{2I,e}^{MITC3(k)} \end{aligned} \quad (18)$$

and  $A_{ES}$ ,  $N_{ES}$  and  $E_{ES}$  are corresponding to the area, the number of nodes and elements of smoothing domains.

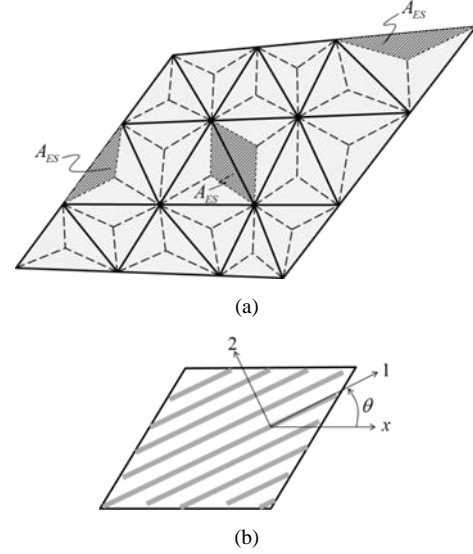


Fig. 2. (a) Edge-based smoothing domains for a mesh of the ES-MITC3 elements; (b) Direction definition of the orthotropic layer  $\langle k \rangle$  in the cartesian coordinate system

Consider layer  $\langle k \rangle$  with the orthotropic material properties  $E_1^{\langle k \rangle}$ ,  $E_2^{\langle k \rangle}$ ,  $G_{12}^{\langle k \rangle}$ ,  $G_{13}^{\langle k \rangle}$ ,  $G_{23}^{\langle k \rangle}$ ,  $\nu_{12}^{\langle k \rangle}$  and  $\nu_{21}^{\langle k \rangle}$ , and  $\theta_k$  is the angle between the  $E_1^{\langle k \rangle}$ -direction and  $x$ -axis as in Fig. 2b. The stress-strain relationships of layer  $\langle k \rangle$  are [1]

$$\begin{Bmatrix} \boldsymbol{\sigma}_m^{(k)} \\ \boldsymbol{\tau}_s^{(k)} \end{Bmatrix} = \begin{bmatrix} Q_{xx}^{(k)} & Q_{xy}^{(k)} & Q_{xp}^{(k)} & 0 & 0 \\ Q_{yx}^{(k)} & Q_{yy}^{(k)} & Q_{yp}^{(k)} & 0 & 0 \\ Q_{px}^{(k)} & Q_{py}^{(k)} & Q_{pp}^{(k)} & 0 & 0 \\ 0 & 0 & 0 & Q_{nm}^{(k)} & Q_{nn}^{(k)} \\ 0 & 0 & 0 & Q_{mn}^{(k)} & Q_{mm}^{(k)} \end{bmatrix} \begin{Bmatrix} \tilde{\boldsymbol{\varepsilon}}^{(k)} \\ \tilde{\boldsymbol{\gamma}}^{MITC3(k)} \end{Bmatrix} \quad (19)$$

in which

$$\begin{aligned} Q_{xx}^{(k)} &= Q_{11}^{(k)} \cos^4 \theta_k + Q_{22}^{(k)} \sin^4 \theta_k + 2(Q_{12}^{(k)} + 2Q_{66}^{(k)}) \sin^2 \theta_k \cos^2 \theta_k \\ Q_{yy}^{(k)} &= Q_{11}^{(k)} \sin^4 \theta_k + Q_{22}^{(k)} \cos^4 \theta_k + 2(Q_{12}^{(k)} + 2Q_{66}^{(k)}) \sin^2 \theta_k \cos^2 \theta_k \\ Q_{pp}^{(k)} &= (Q_{11}^{(k)} + Q_{22}^{(k)} - 2Q_{12}^{(k)} - 2Q_{66}^{(k)}) \sin^2 \theta_k \cos^2 \theta_k + Q_{66}^{(k)} (\sin^4 \theta_k + \cos^4 \theta_k) \\ Q_{xy}^{(k)} &= Q_{12}^{(k)} = (Q_{11}^{(k)} + Q_{22}^{(k)} - 4Q_{66}^{(k)}) \sin^2 \theta_k \cos^2 \theta_k + Q_{12}^{(k)} (\sin^4 \theta_k + \cos^4 \theta_k) \\ Q_{nm}^{(k)} &= Q_{44}^{(k)} \cos^2 \theta_k + Q_{55}^{(k)} \sin^2 \theta_k \\ Q_{mn}^{(k)} &= Q_{55}^{(k)} \cos^2 \theta_k + Q_{44}^{(k)} \sin^2 \theta_k; Q_{nm}^{(k)} = Q_{mn}^{(k)} = (Q_{55}^{(k)} - Q_{44}^{(k)}) \sin \theta_k \cos \theta_k \end{aligned} \quad (20)$$

$$Q_{xp}^{(k)} = Q_{px}^{(k)} = (Q_{11}^{(k)} - Q_{12}^{(k)} - 2Q_{66}^{(k)}) \sin \theta_k \cos^3 \theta_k + (Q_{12}^{(k)} - Q_{22}^{(k)} + 2Q_{66}^{(k)}) \sin^3 \theta_k \cos \theta_k$$

$$Q_{yp}^{(k)} = Q_{py}^{(k)} = (Q_{11}^{(k)} - Q_{12}^{(k)} - 2Q_{66}^{(k)}) \sin^3 \theta_k \cos \theta_k + (Q_{12}^{(k)} - Q_{22}^{(k)} + 2Q_{66}^{(k)}) \sin \theta_k \cos^3 \theta_k$$

$$\text{with } Q_{11}^{(k)} = \frac{E_1^{(k)}}{1 - \nu_{12}^{(k)} \nu_{21}^{(k)}}; Q_{22}^{(k)} = \frac{E_2^{(k)}}{1 - \nu_{12}^{(k)} \nu_{21}^{(k)}}; Q_{12}^{(k)} = Q_{21}^{(k)} = \frac{\nu_{12}^{(k)} E_2^{(k)}}{1 - \nu_{12}^{(k)} \nu_{21}^{(k)}};$$

$$Q_{66}^{(k)} = G_{12}^{(k)}; Q_{44}^{(k)} = G_{23}^{(k)}; Q_{55}^{(k)} = G_{13}^{(k)}.$$

The principle of virtual work of the 3-layer sandwich plate with the area  $A_0$  under transverse distributed load  $q$  is expressed by

$$\sum_{k=1}^3 \int_{A_0} (\delta \tilde{\mathbf{e}}^{(k)})^T \tilde{\mathbf{\sigma}}_m^{(k)} dA + \sum_{k=1}^3 \int_{A_0} (\delta \tilde{\mathbf{Y}}^{MITC3(k)})^T \tilde{\mathbf{\tau}}_s^{(k)} dA = \int_{A_0} \delta w_0 q dA \quad (21)$$

Substituting the relationships between the stresses and strains in Eq. (19), and the strains and nodal displacements in Eq. (17) into Eq. (21), the discretized equilibrium equations can be obtained  $\mathbf{Kd} = \mathbf{F}$ , in which  $\mathbf{d}$  is the nodal displacements of the plate,  $\mathbf{K}$  and  $\mathbf{F}$  are respectively the plate stiffness matrix and force vector assembled from

$$\mathbf{k}_U = \sum_{k=1}^3 (\tilde{\mathbf{B}}_I^{(k)})^T \mathbf{D}_b^{(k)} \tilde{\mathbf{B}}_I^{(k)} A_{ES} + \sum_{k=1}^3 (\tilde{\mathbf{S}}_I^{(k)})^T \mathbf{D}_s^{(k)} \tilde{\mathbf{S}}_I^{(k)} A_{ES} \quad (22)$$

$$\mathbf{f}_I = \int_{A_0} [N_I \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]^T q dA$$

with  $\tilde{\mathbf{B}}_I^{(k)} = [\tilde{\mathbf{B}}_{1I}^{(k)} \ \tilde{\mathbf{B}}_{2I}^{(k)} \ \tilde{\mathbf{B}}_{3I}^{(k)} \ \tilde{\mathbf{B}}_{4I}^{(k)}]$ ;  $\tilde{\mathbf{S}}_I^{(k)} = [\tilde{\mathbf{S}}_{1I}^{(k)} \ \tilde{\mathbf{S}}_{2I}^{(k)}]$  and

$$\mathbf{D}_b^{(k)} = \begin{bmatrix} \mathbf{D}_{b2}^{(k)} & \mathbf{D}_{b1}^{(k)} & \mathbf{D}_{b1}^{(k)} & \mathbf{D}_{b4}^{(k)} \\ \mathbf{D}_{b1}^{(k)} & \mathbf{D}_{b0}^{(k)} & \mathbf{D}_{b0}^{(k)} & \mathbf{D}_{b3}^{(k)} \\ \mathbf{D}_{b1}^{(k)} & \mathbf{D}_{b0}^{(k)} & \mathbf{D}_{b0}^{(k)} & \mathbf{D}_{b3}^{(k)} \\ \mathbf{D}_{b4}^{(k)} & \mathbf{D}_{b3}^{(k)} & \mathbf{D}_{b3}^{(k)} & \mathbf{D}_{b6}^{(k)} \end{bmatrix}; \mathbf{D}_s^{(k)} = \begin{bmatrix} \mathbf{D}_{s0}^{(k)} & \mathbf{D}_{s2}^{(k)} \\ \mathbf{D}_{s2}^{(k)} & \mathbf{D}_{s4}^{(k)} \end{bmatrix} \quad (23)$$

herein,

$$(D_{b0,ij}^{(k)}, D_{b1,ij}^{(k)}, D_{b2,ij}^{(k)}, D_{b3,ij}^{(k)}, D_{b4,ij}^{(k)}, D_{b6,ij}^{(k)}) = \int_{-t_i/2}^{t_i/2} (1, z_k, z_k^2, z_k^3, z_k^4, z_k^6) Q_{ij}^{(k)} dz \text{ with } i, j = x, y, p$$

$$(D_{s0,ij}^{(k)}, D_{s2,ij}^{(k)}, D_{s4,ij}^{(k)}) = \int_{-t_i/2}^{t_i/2} (1, z_k^2, z_k^4) Q_{ij}^{(k)} dz \text{ with } i, j = m, n$$

### III. NUMERICAL VERIFICATION

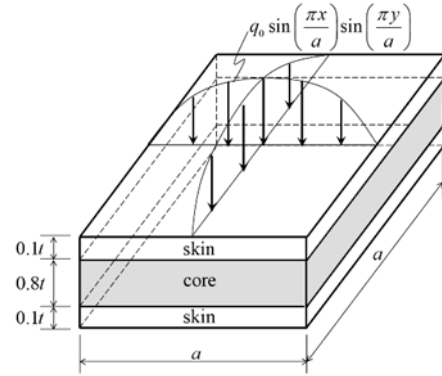
In this section, the efficiency of the proposed element is verified by static analysis of simply supported square sandwich plates under sinusoidal or uniformly distributed loads. The square sandwich plates with length  $a$  and thickness  $t$  consist of skin, core and skin with the corresponding thicknesses  $0.1t$ ,  $0.8t$  and  $0.1t$  as shown in Fig. 3.

#### A. Sandwich Plate under Sinusoidal Distributed Load

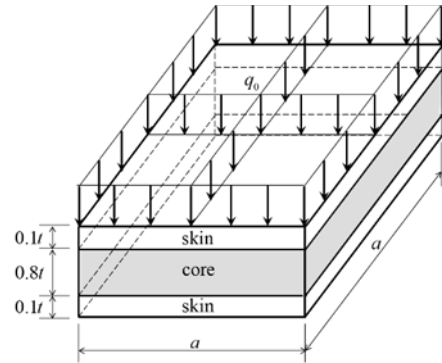
Consider the square sandwich plate subjected to sinusoidal distributed load  $q_0 \sin(\pi x/a) \sin(\pi y/a)$  as depicted in Fig. 3a. The material properties of the skins are  $E_2^{<s>} = 1$ ,  $E_1^{<s>} = 25E_2^{<s>}$ ,  $G_{12}^{<s>} = G_{13}^{<s>} = 0.5E_2^{<s>}$ ,  $G_{23}^{<s>} = 0.2E_2^{<s>}$ ,  $\nu_{12}^{<s>} = 0.25$ , and  $E_2^{<c>} = 0.04$ ,  $E_1^{<c>} = E_2^{<c>}$ ,  $G_{12}^{<c>} = 0.016$ ,  $G_{13}^{<c>} = G_{23}^{<c>} = 0.06$ ,  $\nu_{12}^{<c>} = 0.25$  for the core.

The sandwich plate is discretized by meshes of  $N$  ES-MITC3 elements on each edge to investigate convergence of the deflection at the plate's center. As compared with the elasticity solution [15], relative errors of the deflections given

by the different meshes of  $N = 8, 12, 16, 20, 24$  ES-MITC3 elements are demonstrated in Fig. 4 for  $a/t = 4$  and  $a/t = 100$ . With the mesh  $N = 24$ , the ES-MITC3 element predicts the deflections with relative errors below 1%. Fig. 4 indicates that the accuracy and convergence of the suggested approach is superior to those of the ES-MITC3 element based on the FSD-LW theory [11], especially in the thin plate  $a/t = 100$ .

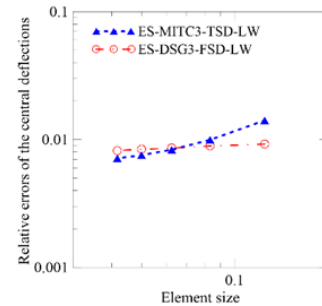


(a) Sinusoidal distributed load

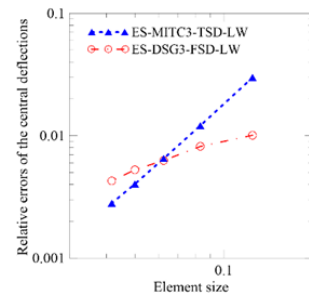


(b) Uniformly distributed load

Fig. 3. Geometry and sinusoidal or uniformly distributed loads of simply supported square sandwich plates



(a)  $a/t = 4$



(b)  $a/t = 100$

Fig. 4. Convergence rates of the deflections at the center of simply supported square sandwich plates subjected to sinusoidal distributed load

TABLE I. NORMALIZED DEFLECTIONS AND STRESSES OF SIMPLY SUPPORTED SQUARE SANDWICH PLATES SUBJECTED TO SINUSOIDAL DISTRIBUTED LOAD

$a/t$	Methods	$\bar{w}$	$\bar{\sigma}_{xx}$	$\bar{\sigma}_{yy}$	$\bar{\tau}_{xz}$	$\bar{\tau}_{yz}$
4	Elasticity [15]	7.5962	1.5560	0.2595	0.2390	0.1072
	Q18-3D-FEM-LW [7]	-	1.5700	0.2600	0.2300	0.1080
	Q9-FEM-HOZT [6]	7.5822	1.5306	0.2581	0.2436	0.1147
	ES-DSG3-FSD-LW [11]	7.6585	1.4624	0.2484	0.2352	0.1025
	MITC3-TSD-LW	7.6201	1.5037	0.2487	0.2372	0.1052
	ES-MITC3-TSDT [16]	7.1564	1.4959	0.2377	0.2816	0.1166
	ES-MITC3-TSD-LW	7.6503	1.5096	0.2495	0.2386	0.1062
10	Elasticity [15]	2.2004	1.1153	0.1104	0.300	0.0527
	Q18-3D-FEM-LW [7]	-	1.1590	0.1110	0.3030	0.0550
	Q9-FEM-HOZT [6]	2.1775	1.1528	0.1143	0.3058	0.0575
	ES-DSG3-FSD-LW [11]	2.1991	1.1407	0.1083	0.2965	0.0506
	MITC3-TSD-LW	2.1922	1.1390	0.1081	0.2973	0.0519
	ES-MITC3-TSD-LW	2.1903	1.1436	0.1085	0.2988	0.0528
	ES-MITC3-TSD-LW	2.1903	1.1436	0.1085	0.2988	0.0528
20	Elasticity [15]	1.2264	1.1100	0.0700	0.3170	0.0361
	Q18-3D-FEM-LW [7]	-	1.1100	0.0700	0.3170	0.0360
	Q9-FEM-HOZT [6]	1.2121	1.1103	0.0742	0.3272	0.0399
	ES-DSG3-FSD-LW [11]	1.2228	1.1017	0.0692	0.3147	0.0351
	MITC3-TSD-LW	1.2195	1.0983	0.0691	0.3145	0.0356
	ES-MITC3-TSDT [16]	1.1906	1.1019	0.0678	0.3666	0.0409
	ES-MITC3-TSD-LW	1.2241	1.1028	0.0693	0.3161	0.0365
100	Elasticity [15]	0.8923	1.0980	0.0550	0.3240	0.0297
	Q9-FEM-HOZT [6]	0.8814	1.0982	0.0592	0.3426	0.0322
	ES-DSG3-FSD-LW [11]	0.8885	1.0904	0.0546	0.3241	0.0316
	MITC3-TSD-LW	0.8863	1.0868	0.0544	0.3193	0.0277
	ES-MITC3-TSDT [16]	0.8876	1.0902	0.0545	0.3719	0.0333
	ES-MITC3-TSD-LW	0.8898	1.0912	0.0546	0.3212	0.0289
	ES-MITC3-TSD-LW	0.8898	1.0912	0.0546	0.3212	0.0289

The normalized deflections and stresses

$$\begin{aligned}\bar{w}_c &= \frac{100E_2t^3}{q_0a^4} w(a/2, a/2); \\ \bar{\sigma}_{xx} &= \frac{t^2}{q_0a^2} \sigma_{xx}(a/2, a/2, t/2); \quad \bar{\sigma}_{yy} = \frac{t^2}{q_0a^2} \sigma_{yy}(a/2, a/2, t/2) \quad (24) \\ \bar{\tau}_{xz} &= \frac{t}{q_0a} \tau_{xz}(0, a/2, 0); \quad \bar{\tau}_{yz} = \frac{t}{q_0a} \tau_{yz}(a/2, 0, 0)\end{aligned}$$

obtained by the mesh of  $N = 24$  ES-MITC3 elements are listed in Table 1. With various values of  $a/t = 4, 10, 20$  and  $100$ , in comparison with the elasticity solution [15], the presented element usually gives better results than those provided by the ES-DSG3 based on the FSD-LW theory [11] and MITC3 based on the TSD-LW theory because the presented element employs both the TSD-LW theory and the edge-based strain smoothing method. The results of the ES-MITC3 element also well agree with those of the 3-dimensional quadrilateral element Q18 using layerwise theory (Q18-3D-FEM-LW) [7] and the quadrilateral element Q9 employing the higher-order zig-zag theory (Q9-FEM-HOZT) [6]. Because of using the LW theory, the proposed element predicts more accurate deflection and stresses than those of the ES-MITC3 element employing the ESL theory, especially for thick plates. However, the presented element consumes the computational time about 3.5 times those using the ESL theory.

#### B. Sandwich Plate under Uniformly Distributed Load

In this example, the sandwich plate with  $a/t = 10$  subjected to uniformly distributed load  $q_0$  is studied as illustrated in Fig.

3b. The material's constitutive matrices of the core and the skins respectively are  $\mathbf{Q}_{core}$  and  $\mathbf{Q}_{skin} = R\mathbf{Q}_{core}$ , in which

$$\mathbf{Q}_{core} = \begin{bmatrix} 0.999781 & 0.231192 & 0 & 0 & 0 \\ 0.231192 & 0.524886 & 0 & 0 & 0 \\ 0 & 0 & 0.262931 & 0 & 0 \\ 0 & 0 & 0 & 0.266810 & 0 \\ 0 & 0 & 0 & 0 & 0.159914 \end{bmatrix} \quad (25)$$

To investigate convergence of the presented element in this example, the deflections at the plate's center determined by the meshes of  $N = 8, 12, 16, 20$  and  $24$  are compared with those provided by the analytical method [17] for  $R = 5$  and  $15$ . The graphs in logarithmic scale of the mesh sizes versus the relative errors given by the ES-MITC3 and CS-DSG3 elements are plotted in Fig. 5. For both cases  $R = 5$  and  $15$ , the ES-MITC3 element achieves better accuracy and convergence of the deflections than the CS-DGS3 element does [12]. The mesh of  $N = 24$  also gives the convergent deflections with the relative errors below 0.2% and is used to predict the normalized deflections and stresses as presented in Table 2.

$$\begin{aligned}\bar{w}_c &= \frac{0.999781}{q_0t} w(a/2, a/2); \quad \bar{\sigma}_{xx} = \frac{\sigma_{xx}(a/2, a/2, t/2)}{q_0} \\ \bar{\sigma}_{yy} &= \frac{\sigma_{yy}(a/2, a/2, 2t/5)}{q_0}; \quad \bar{\sigma}_{xx}^* = \frac{\sigma_{xx}(a/2, a/2, 2t/5)}{q_0} \quad (26) \\ \bar{\tau}_{xz} &= \frac{\tau_{xz}(0, a/2, 0)}{q_0}\end{aligned}$$

Owing to the smoothing strain method, the ES-MITC3 element obtains larger and more accurate values of the

deflections and stresses than those given by the MITC3 element. The static behaviors provided by the presented element, the 4-node quadrilateral element based on the trigonometric shear deformation LW theory (Q4-FEM-TrSD-

LW) [5] and the FSD-LW theory-typed isogeometric analysis (IGA-FSD-LW) [4] are similar for the cases of  $R = 5, 10$  and  $15$ .

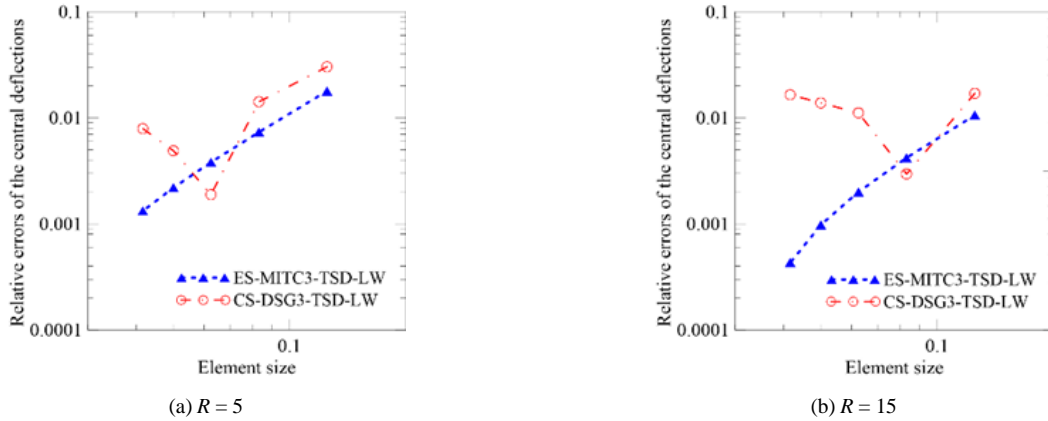


Fig. 5. Convergence rates of the deflections at the center of simply supported square sandwich plates subjected to uniformly distributed load

TABLE II. NORMALIZED DEFLECTIONS AND STRESSES OF SIMPLY SUPPORTED SQUARE SANDWICH PLATES SUBJECTED TO UNIFORMLY DISTRIBUTED LOAD

$R$	Methods	$\bar{w}$	$\bar{\sigma}_{xx}$	$\bar{\sigma}_{yy}$	$\bar{\sigma}_{xz}^*$	$\bar{\tau}_{xz}$
5	Analytics [17]	258.9700	60.3530	30.0970	9.3400	4.3641
	IGA-FSD-LW [4]	258.8347	60.2535	30.1094	9.3020	-
	Q4-FEM-TrSD-LW [5]	256.7060	60.5250	30.1770	9.4120	-
	CS-DSG3-TSD-LW [12]	261.0157	60.3499	30.2236	9.3010	4.0434
	MITC3-TSD-LW	257.9144	59.9463	29.9783	9.2535	4.2379
	ES-MITC3-TSD-LW	258.6256	60.0744	30.0349	9.2734	4.2941
10	Analytics [17]	152.3300	64.6500	33.9700	5.1310	3.1470
	IGA-FSD-LW [4]	159.4059	65.2297	33.5008	4.8733	-
	Q4-FEM-TrSD-LW [5]	155.4980	65.5420	33.5910	4.9710	-
	CS-DSG3-TSD-LW [12]	161.4662	65.4388	33.6964	4.8756	3.9375
	MITC3-TSD-LW	158.7989	64.8684	33.3381	4.845	3.9817
	ES-MITC3-TSD-LW	159.2461	65.0049	33.401	4.8554	4.0332
15	Analytics [17]	121.7200	66.7870	34.9550	3.2380	3.9638
	IGA-FSD-LW [4]	121.7764	66.6792	35.0797	3.2111	-
	Q4-FEM-TrSD-LW [5]	115.9190	67.1850	35.0810	3.3180	-
	CS-DSG3-TSD-LW [12]	123.7309	66.9754	35.3273	3.2135	3.8623
	MITC3-TSD-LW	121.3194	66.3008	34.9017	3.1916	3.8561
	ES-MITC3-TSD-LW	121.6675	66.4372	34.9685	3.1984	3.9052

#### IV. CONCLUSIONS

A formulation of the ES-MITC3 element for static analysis of sandwich plates according to the TSD-LW theory has been presented. The deflections and stresses of the sandwich plates with various length-to-thickness ratios under sinusoidal or uniformly distributed loads were predicted by the presented element. Numerical results showed that the ES-MITC3 element overcame the shear-locking phenomenon and was more accurate than the MITC3 element. Using the same TSD-LW theory, the ES-MITC3 element obtained better deflection convergences than the CS-DSG3 element did. The suggested approach based on the LW theory predicted better static behaviors but consumed larger amount of computational time than those using the ESL theory. The proposed approach in this study also provided similar results to those obtained by IGA or such higher-order elements as Q4, Q9 or Q18 ones.

#### REFERENCES

- [1] J. N. Reddy, *Mechanics of Laminated Composite Plates and Shells - Theory and Analysis*, Second. CRC Press, 2004.
- [2] K. M. Liew, Z. Z. Pan, and L. W. Zhang, "An overview of layerwise theories for composite laminates and structures: Development, numerical implementation and application," *Composite Structures*, vol. 216, pp. 240–259, 2019.
- [3] A. J. M. Ferreira, "Analysis of Composite Plates Using a Layerwise Theory and Multiquadrics Discretization," *Mechanics of Advanced Materials and Structures*, vol. 12, no. 2, pp. 99–112, 2005.
- [4] C. H. Thai, A. J. M. Ferreira, E. Carrera, and H. Nguyen-Xuan, "Isogeometric analysis of laminated composite and sandwich plates using a layerwise deformation theory," *Composite Structures*, vol. 104, pp. 196–214, 2013.
- [5] J. L. Mantari, A. S. Oktem, and C. Guedes Soares, "A new trigonometric layerwise shear deformation theory for the finite element analysis of laminated composite and sandwich plates," *Computers & Structures*, vol. 94–95, pp. 45–53, 2012.
- [6] H. D. Chalak, A. Chakrabarti, Mohd. A. Iqbal, and A. Hamid Sheikh, "An improved  $C^0$  FE model for the analysis of laminated sandwich

- plate with soft core,” *Finite Elements in Analysis and Design*, vol. 56, pp. 20–31, 2012.
- [7] G. S. Ramtekkar, Y. M. Desai, and A. H. Shah, “Application of a three-dimensional mixed finite element model to the flexure of sandwich plate,” *Computers & Structures*, vol. 81, no. 22, pp. 2183–2198, 2003.
- [8] K.-U. Bletzinger, M. Bischoff, and E. Ramm, “A unified approach for shear-locking-free triangular and rectangular shell finite elements,” *Computers & Structures*, vol. 75, no. 3, pp. 321–334, 2000.
- [9] P.-S. Lee and K.-J. Bathe, “Development of MITC isotropic triangular shell finite elements,” *Computers & Structures*, vol. 82, no. 11–12, pp. 945–962, 2004.
- [10] G. R. Liu and T. Nguyen-Thoi, *Smoothed Finite Element Methods*. CRC Press, 2010.
- [11] P. Phung-Van, C. H. Thai, T. Nguyen-Thoi, and H. Nguyen-Xuan, “Static and free vibration analyses of composite and sandwich plates by an edge-based smoothed discrete shear gap method (ES-DSG3) using triangular elements based on layerwise theory,” *Composites Part B: Engineering*, vol. 60, pp. 227–238, 2014.
- [12] P. Phung-Van, T. Nguyen-Thoi, H. Dang-Trung, and N. Nguyen-Minh, “A cell-based smoothed discrete shear gap method (CS-FEM-DSG3) using layerwise theory based on the C0-HSDT for analyses of composite plates,” *Composite Structures*, vol. 111, pp. 553–565, 2014.
- [13] T. Chau-Dinh, “Static analysis of laminated composite plates based on a layerwise model using ES-MITC3 elements,” *Review of Ministry of Construction*, vol. 8/2017, pp. 75–82, 2017.
- [14] T. Chau-Dinh, Q. Nguyen-Duy, and H. Nguyen-Xuan, “Improvement on MITC3 plate finite element using edge-based strain smoothing enhancement for plate analysis,” *Acta Mech*, vol. 228, no. 6, pp. 2141–2163, 2017.
- [15] N. J. Pagano, “Exact Solutions for Rectangular Bidirectional Composites and Sandwich Plates,” *Journal of Composite Materials*, vol. 4, no. 1, pp. 20–34, 1970.
- [16] T. Chau-Dinh, H. Nguyen-Van, and H. Nguyen-Van, “Static analysis of functionally graded plates using the high-order shear deformation theory by MITC3 plate elements having strains smoothed on edges,” in *Design, Manufacturing and Applications of Composites*, Ho Chi Minh City, Vietnam, Aug. 2016, pp. 252–264.
- [17] S. Srinivas, “A refined analysis of composite laminates,” *Journal of Sound and Vibration*, vol. 30, no. 4, pp. 495–507, 1973.



# P2C2-Popular Content Prediction and Collaboration in mobile edge caching

Dung Ong Mau  
Faculty of Electronics Technology  
Industrial University of HCMC.  
Ho Chi Minh, VietNam  
ongmaudung@iuh.edu.vn

Anh Phan Tuan  
Faculty of Electronics Technology  
Industrial University of HCMC.  
Ho Chi Minh, VietNam  
phantuananh@iuh.edu.vn

Tuyen Dinh Quang  
Faculty of Electronics Technology  
Industrial University of HCMC.  
Ho Chi Minh, VietNam  
dinhquangtuyen@iuh.edu.vn

Quyen Le Ly Quyen  
Faculty of Electronics Technology  
Industrial University of HCMC.  
Ho Chi Minh, VietNam  
lelyquyenquyen@iuh.edu.vn

Nga Vu Thi Hong  
Faculty of Electronics Technology  
Industrial University of HCMC.  
Ho Chi Minh, VietNam  
vuthihongnga@iuh.edu.vn

**Abstract**—Evolution from today's host-centric network architecture to a data-centric network architecture, named data networking (NDN) is a proposed Future Internet architecture based on content name-oriented that pushes and caches data to edge gateways/routers. If the content is popular, the previous queried content can be reused multiple times and it should be kept in the limited size of storage longer than unpopular contents. In the context of mobile data networks, researchers conclude that NDN is one of the best fit solutions in mobile edge caching. Thus, it is critical to study an efficient replacement policy to achieve a higher hitting rate than the original policies applied in NDN protocol. This paper proposes a novel Popular Content Prediction and Collaboration (P2C2) in the context of mobile edge caching. OPNET simulation tool is used to conduct long term evolution (LTE) network integrated with NDN protocol and our proposed P2C2 replacement policy. The simulation results show that the hitting rate value of P2C2 is higher than Time-aware Least Recently Used (TLRU) and Least Frequently Used (LFU) significantly, and P2C2 reduces the traffic of back-haul while facilitating the offloading of server traffic.

**Index Terms**—Replacement policy, named-data network, mobile edge network, green information technology

## I. INTRODUCTION

With the convergence of cloud computing with social media and mobile communication, the types of data traffic are becoming more diverse while the number of Internet users is increasing exponentially [1]. Millions of multimedia files are generated and shared by producers and consumers, which poses high requirements for the network bandwidth and data storage, and causes overload on the server [2] [3]. Thus, Internet users and mobile subscribers often feel unsatisfied with the performance in terms of delay, jitter and throughput.

Under such a background of traffic growing and number of massive uses increasing, existing network architecture meets challenges. The network technology for today's Internet was created in the 1960s and until this time, it is still based on host-centric network architecture and the peer-to-peer (P2P) principle. On the other hand, in the future of social media,

Internet subscribers care about the data content they wish to get much more than where the content comes from [4].

Evolution from information centric networking (ICN), named data networking (NDN) was first proposed by V. Jacobson in 2009 and at this time, NDN is not only a theory proposal but also capability for the real world implementation [5]- [8]. In order to alleviate the bandwidth problem while considering the feature of Internet subscribers, the NDN is proposed to effectively distribute popular data content to a huge number of users [9]. By pushing and caching popular arrived content as long as possible to edge gateways/routers, the NDN trends to spread caches all over the network and maximize the probability of sharing with minimal upstream bandwidth demand and lowest downstream latency. For this reason, NDN nodes request and receive content one time, and arrival content can be consumed many times. When a huge number of users request for the same data content, the NDN gains more effective distribution content than P2P connectivity. To prevent buffers from overflowing, NDN node maintains a timer to track time-to-live (TTL) for each content and does a replacement algorithm rather than holds all contents permanently [11]. Time-aware Least Recently Used (TLRU) and Least Frequently Used (LFU) replacement policies are two famous algorithms for ICN but they ignore the advantage of NDN as follows.

- TLRU and LFU only base on property of object name to make replacement decisions. Since, TLRU uses a time-stamp on object name while LFU uses a frequency occurrence of object name.
- TLRU and LFU cannot recognize the popularity ranking of new content coming and cause highly accessed content to be replaced by the new one.
- All contents have the same lifetime while their popularity ranking are very skew.

In order to fully exhibit better performance of NDN com-

pared to traditional network architectures, it's critical to design a highly efficient replacement policy for the NDN. In this paper, we propose a novel replacement policy, named Popular Content Prediction and Collaboration (P2C2) in mobile edge caching. The P2C2 considers multi-level prefix (MP) of all contents in the CS, then determines popularity ranking and gives suitable lifetime for each content, including new coming content. Moreover P2C2 periodically exchanges MP tables among close by NDN nodes to improve the performance of cache.

Due to the important as compared to the existing TLRU or LFU solutions, the P2C2 schemes have the following unique features:

- We carefully investigate the characteristics of NDN where the name of data content includes multi-levels of prefixes. We add the MP table to handle multi-level prefix names with timer and counters.
- The way Internet users request content is fine turned with Pareto principle, that is 20% popular content is requested by 80% number of users.
- The prefix of popular data contents always appears with a high probability and vice versa.
- Content with high popular prefix (e.g. popular publisher server, related novel content) will be popular too.

From the background of NDN, we conduct NDN protocol on top of IP layer in the context of mobile edge caching for LTE networks. The existing TLRU, LFU and our proposal P2C2 replacement policy have been successfully constructed in the NDN. Our simulation results prove that NDN is a good solution for existing challenges of traditional IP networks. And P2C2 outperforms TLRU, LFU with highly effective caching.

The remainder of this paper is organized as follows. We review some new proposal effective caching policies for NDN in Section II. In Section III, we present our replacement algorithm. Section IV describes our network architecture simulation and verifies simulation results. Finally, Section V concludes this paper.

## II. RELATED WORKS

As a proposed future Internet architecture, Named Data Networking (NDN) is designed to network the world of computing devices by naming data instead of data containers as IP does today [10]. In NDN node, two types of packets are interest packet (IntPk) and data packet (DataPk) identified data content by name, and three data structures are Forwarding Information Base (FIB), Pending Interest Table (PIT) and Content Store (CS). When an NDN node receives the IntPk, it looks for data content in CS and DataPk is replied if found content, otherwise IntPk is checked in PIT. The PIT keeps track of list of users for unsatisfied IntPk while the FIB is a table of outbound faces for IntPk. And unsatisfied IntPk are forwarded upstream toward potential content sources. Then returned DataPk will be sent to single or multi-downstream and CS keeps received content for the next response. For this reason, NDN can naturally reflect the popularity, but we need an acceleration.

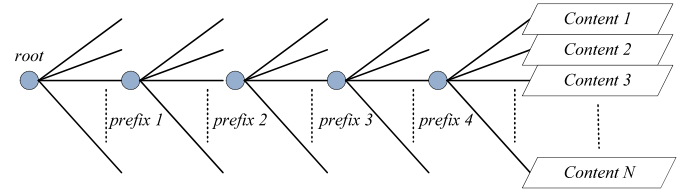


Fig. 1. An example prefix tree structure.

In the most recent article [12], some of potential solutions of content-centric networking (CCN) known as NDN based edge caching for 5G network have been proposed. The edge caching performance gains the QoS of end-users as well as traffic offloading significantly.

Similar to [12], there is a lot of research focusing on mobile edge caching with content popularity prediction, typically in [13] and [14]. In [13], by analyzing user preferences, authors propose algorithms that perform different caching which achieve better performance than general methods. And in [14] artificial neural networks are used to learn the content popularity matrix by observing the instantaneous demands over time.

Last but not least, because of the limits on storage size, limits on the ability to process data at the nodes, simple and effective caching at NDN nodes should be taken into consideration. Different from previous works, OPNET Modeler is used to conduct our proposed protocol in the mobile edge caching network. Indeed, a distinguished OPNET feature is the detailed models of network equipment and everything in OPNET happens in a way that closely follows real network behavior [15]. Especially, packet processing above the data layer in OPNET closely mirrors the way real-world network equipment works. We are going to present our proposal P2C2 replacement policy in the next section.

## III. P2C2 REPLACEMENT ALGORITHM

In replacement policy, it is not enough to consider only the name of content. That is to say, well-known publishers always make popular content and the history of prefix can tell us everything. We handle the history of all multi-level prefixes by using an MP table, then effective caching is done with extending lifetime for popular content. The other important thing is collaboration between neighboring edge routers. Indeed, if neighbor routers exchange the MP table, they will gain effective caching.

### A. Multi-level Prefix structure

Names of content are hierarchically structured and human readable. NDN node uses the longest prefix mapping to lookup the name in CS, PIT and FIB when it serves for arrival IntPk. For notational convenience, names are presented like Uniform Resource Identifier (URI) with “/” characters separating components. Fig. 1 is an example structure of name with 5-levels tree form. On the top of the tree structure, “root” element is called globally routable name [10]. NDN published

TABLE I  
SUM UP MULTIPLE SKEWNESS FACTORS OF POPULAR CONTENT.

Factor $\theta\%$	Prefix index $id(1) \rightarrow id(4)$	Content index( $N$ ) (Number of index with Probability)	Number of popular contents over total contents
90	Uniform(1, 5)	3 with 94% 47 with 6%	$(3525/31250) \simeq 11\%$
80	Uniform(1, 5)	6 with 88% 44 with 12%	$(6600/31250) \simeq 21\%$
70	Uniform(1, 5)	10 with 80% 40 with 20%	$(10000/31250) \simeq 32\%$
60	Uniform(1, 5)	14 with 72% 36 with 28%	$(12600/31250) \simeq 40\%$
50	Uniform(1, 5)	50 with 50%	$(15625/31250) = 50\%$

servers repeat at a time interval to broadcast the roots name to the whole network. When NDN nodes receive the roots name, they add to their routing table (FIB) and then continue forward roots name to neighbor nodes. In some cases, either NDN published server or NDN nodes at the middle of network are disconnected, all NDN nodes can realize a changing of network and reconfigure FIB by receiving periodic roots name.

Power-law distribution supported by OPNET modeler is used to create a huge number of prefix indexes. For instance, a large number of names are based on the tree structure “ $root/id(1)/id(2)/\dots/id(i)/\dots/id(M)$ ”, with  $id(i)$  follows the probability density function (PDF) in equation (1) and  $1 \leq i \leq M$ .

$$f(x) = \frac{c \cdot x^{c-1}}{b^c} \quad (1)$$

With  $x$  is a continuous random variable and  $0 \leq x \leq b$ ,  $c$  is a shape parameter and  $b$  is a scale parameter. To reflect the real world information, D. Rossi et. al., [16] summarize some of the most relevant system parameters used in related works. In [16], the number of content in the considered catalogs can be as low as 250 objects, topping to 20000 objects. We realize that the object size is extremely small compared to the real world of Internet catalog (e.g.  $10^8$  contents for YouTube and  $5 \times 10^6$  contents for BitTorrent). In other words, the number of contents is all underestimated because of limited simulation. Compared to [16], we expect 30000 contents are good enough for the replacement algorithm performance coverage to final state. For this reason, 5 different prefix indexes are set up from  $id(1)$  to  $id(4)$ , while 50 different indexes are chosen for  $Content(N = 50)$ . If prefix indexes of all levels are selected by Uniform distribution, the total number of names is up to 31250 names exceeding the expected norm.

Let  $0\% \leq \theta \leq 100\%$  is a skewness factor on the number of potential popular content. Recall that  $\theta\%$  of requested contents are focused on  $(100 - \theta)\%$  of popular contents. Table I summarizes a number of popular contents based on a combination of Uniform and power-law distribution.

#### B. Drawback of TLRU and LFU

Normally, all contents in the CS are marked with a timestamp [10]. When contents are responded to satisfy IntPk, the time stamp of used contents will be up-to-date. Moreover, the

CS maintains the same lifetime for all contents and periodically refreshes a memory by checking TTL of all contents. Old contents are deleted if time-out, otherwise recently used content is kept in the memory.

The drawback of TLRU and LFU often happens when memory of the CS is fully filled. Let us consider a situation where memory was in full condition and new unpopular content arrived. Because the CS do not know the popularity ranking of new content, one popular content in CS may be deleted and replaced by a new unpopular one. For this reason, the hitting ratio of TLRU and LFU cannot gain maximum value. By carefully handling the MP table to analyze aspects of timestamp, counter on multi-level prefix name and popularity ranking contents, the P2C2 algorithm is able to predict popularity ranking of new content and avoid to replace high rated content in the memory.

#### C. Popular content ranking and collaboration

We have presented a multi-level prefix, they are hierarchical structure and human-readable. Let  $C_p[i]$  is a counter for the  $i^{th}$  prefix. After the data feedback for the IntPk in the form “ $ccnx://root/id(1)/id(2)/\dots/id(i)/\dots/content(N)$ ”, the CS increases all counters of multi-level prefix matching one time, then lifetime of each content is calculated by equations (2) and (3) as follows.

$$\begin{cases} C_p[1] \leftarrow (root/id(1)) + + \\ C_p[2] \leftarrow (root/id(1)/id(2)) + + \\ \dots\dots\dots \\ C_p[i] \leftarrow (root/id(1)/id(2)/\dots/id(i)) + + \\ \dots\dots\dots \\ C_p[N] \leftarrow (root/id(1)/id(2)/\dots/id(i)/\dots/content(N)) + + \end{cases} \quad (2)$$

$$t_l = t_u \times \sum_{i=1}^N (w[i] \times C_p[i]) \quad (3)$$

With  $t_u$ ,  $w[i]$  are variable lifetime units and weight units of  $C_p[i]$  respectively. Note that unpopular content can utilize vacant memory to contribute a certain percent of hitting rate. For this reason, at the refresh time, if CS still has free space

to store more content, none item is deleted even though the lifetime of some contents could end.

Hitting rate ( $H\%$ ) as shown in equation (4) is an important metric benchmark for replacement algorithm with  $f_{req}$  is a number of arrival  $IntPks$  in one cycle time as known as a request rate of the CS. It should be noted that  $f_{req}$  is equal to sum of number of cache hits and number of cache misses.

$$H(\%) = \frac{\text{Number of cache hits}}{f_{req}} \quad (4)$$

In the mobility environment of Internet mobile subscribers (MSs), the request rate is always an oscillation state with a variable number of users. We handle the effect of dynamic request rate by controlling a lifetime unit ( $t_u$ ). The  $t_u$  value is adaptive adjusted based on measuring a number of requested content in one cycle time (e.g. 60 seconds) as follows.

- When  $f_{req}$  value increases high, the CS is full filled faster and replacement policy has to be implemented. To avoid cache removing some popular contents, we decrease  $t_u$  value associated with the increasing of  $f_{req}$  value. As shown in equation (3), lifetime of all contents ( $t_l$ ) is reduced too. Thus, unpopular contents are found and deleted faster before new contents arrive.
- When  $f_{req}$  value decreases low, we increase  $t_u$  value to keep contents in CS as long as possible. For this reason,  $t_u \times f_{req}$  is a constant.

Among close-by NDN nodes, they can cooperate with others to find out popular contents quickly than itself. The method of operation is very simple, that is all NDN nodes notify their neighbors about the highest popular prefix found in the MP table. In this study, the cooperative interval time is fixed at 30 seconds. If the CS found similar received popular prefixes from others, longer lifetime is given for all relative contents by increasing lifetime unit 10%. For this reason, higher hitting rate and faster convergence speed to final state are achieved.

#### IV. NETWORK SIMULATION AND RESULTS

##### A. Network architecture

OPNET Modeler 16.0 is used to perform NDN protocol with multiple replacement policies in the context of E-UTRAN Node B (eNodeB) caching [18] [19]. NS-3 based NDN (ndnSIM) simulator tool is another famous open source specifically designed for NDN [20]. However, ndnSIM is limited to realistic wireless network simulation while OPNET fully supports WiFi, WiMax and LTE networks.

The LTE network simulation as shown in Fig. 2 includes 7 cells. Each cell has eNodeB, NDN processor node and 25 LTE mobile stations (MSs). There are 2 scenarios in the simulation: (i) the first scenario considers 3 types of replacement algorithm which are stand-alone operation, e.g. TLRU, LFU and P2C-only (Popular Content Prediction), and (ii) the second scenario considers cooperative caching between 7 cells, e.g. P2C2 (Popular Content Prediction and Collaboration).

Through caching policy in ICN has been studied by several researchers [12]- [15], files size ( $F$ ) are random selected by

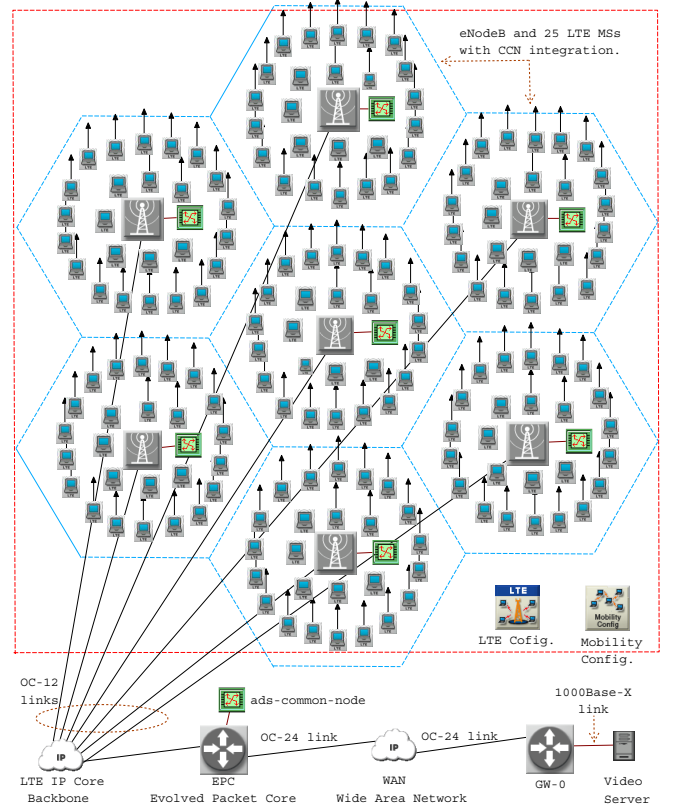


Fig. 2. Network architecture.

Uniform PDF between 1 to 10Mbit(Mb) to accelerate speed simulation. As described in multi-level prefix structure, total consider content ( $|F|$ ) is 31250 with 5 different prefix index. Hence, the average catalog size ( $|F|.F$ ) is approximately 153Gb. Size of memory in the CS ( $C_{size}$ ) is varied at 800Mb, 1.5Gb, 3.0Gb and 4.5Gb to determine the performance caching. The relative cache sizes ( $\frac{C_{size}}{|F|.F}$ ) are taken into account, they are 0.5%, 1.0%, 2.0% and 3.0% respectively. Comparing relative cache sizes with [16] and [17], they are all in range of important value consideration.

Similar with [17], an arrival rate of  $IntPks$  to NDN node is in range from 1 to 10Hz, we config arrival rate varied around 5Hz. With 25 users exist in one cell, users send  $IntPks$  with interval time 5 seconds. Furthermore, the requested contents from users are realistic with randomize starting requested time and slightly random inter-arrival time around 5 seconds for each user. With 3000 seconds simulation, the total  $IntPks$  received by the NDN node is 15000 times which is must higher than popular potential contents. So, hitting rate value can quickly convergence to final state within limited simulation time. Table 2 sums up important parameters for the simulation.

##### B. Simulation results

For better presentation results, we collect and classify all simulation results from OPNET, export them to spreadsheet, and use Matlab to illustrate for final simulation results. Hitting

TABLE II  
SUM UP IMPORTANT PARAMETERS FOR THE SIMULATION

Element	Attribute	Value
LTE	Cell diameter	2000 meters
	Antenna gain	15 dBi
	Maximum transmission power	0.5 W
	Receiver sensitivity	-200 dBm
	eNodeB selection threshold	-110 dBm
	Duplex technique	FDD
	Carrier Frequency (UL/ DL)	1920/ 2110 MHz
	Multiple access (UL/ DL)	SC-FDMA/ OFDMA
	Bandwidth	20 MHz
	Path loss	Free space and without fading
LTE/WAN	Link between enodeBs/gateways Link for server	OC-24 data rate 1000 BaseX
LTE MSs	CCN directory	root/id(1)/id(2)/id(3)/id(4)/content(N)
	File based popularity	As shown in Table I
	Start time	100 + Uniform(0,10)s
	IntPk inter-arrival time	5 + Uniform(0,2)s
	IntPk time-out	2 seconds
Server	CCN root	ccnx://root/
	File size ( $F$ )	Uniform(1,10)Mb
	Number of files ( $ F $ )	31250 files
	Catalog size ( $ F .F$ )	$31250 * E[Uniform(1,10)Mb] \approx 153Gb$
	Packet size	1024 bits
	Publish root's name interval	100 seconds
NDN node	Cache size ( $C_{size}$ )	800Mb; 1.5Gb; 3.0Gb; 4.5Gb
	Relative cache size ( $\frac{C_{size}}{ F .F}$ )	0.5; 1.0; 2.0; 3.0 %
	Replacement policy	TLRU/ LFU/ P2C-only/ P2C2
	Interval refresh CS	1 second
	MP table exchange interval	30 seconds

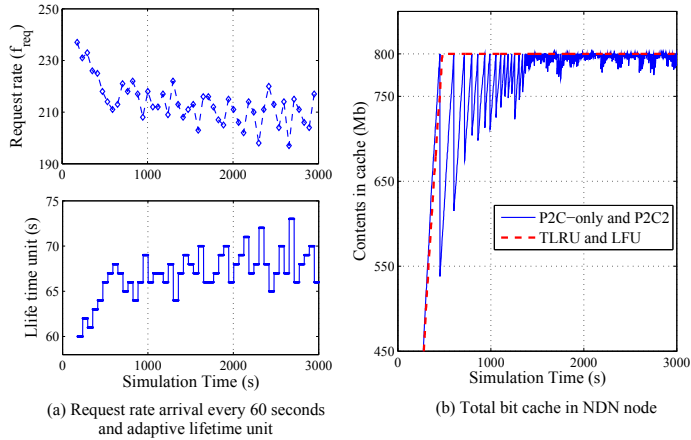


Fig. 3. Impact of request rate on TTL content.

rate at the final coverage state is the most important metric to be verified in the simulation results.

a) *Impact of request rate on TTL contents:* One of the important differences between P2C-only/P2C2 and TLRU/LFU is TTL contents in the CS. For TLRU/LFU, the TTL value of all contents are similar (e.g. 30 minutes) while the TTL for P2C-only/P2C2 is variable due to equation (3). As shown in Fig. 3a, the lifetime unit is adaptive controlled to follow a number of request rate arrival NDN nodes every 60 seconds.

Fig. 3b is a more detailed illustration of the effect of

TTL in the storage area. In case of TLRU/LFU in Fig. 3b, after the total bit in cache reached the up-bound of cache size, the CS is always in the fulfilled condition and leads to the poor replacement of highly accessed content. In case of P2C-only/P2C2, because adaptive TTL is based on popularity ranking contents, some less popular contents are deleted only if the total bit has reached the up-bound of cache size. Then, the total bit storage drops back to a lower value and keeps available space for new coming contents. This feature of P2C-only/P2C2 trends to effective caching.

b) *Impact of replacement policy on hitting ratio:* Fig. 4a illustrates effect of multiple types replacement policy when we fixed relative cache size at 2%. With the same network configuration and the limited size of memory, P2C2 achieves good performance with the highest hitting ratio, followed by P2C-only, TLRU and LFU, respectively.

Fig. 4b shows the performance of multiple replacement policies when the skewness factor is varied from 50% to 90%, and the relative cache size is 2%. As shown in Fig. 4b, when the skewness factor is near 90%, a small amount of content attracts large numbers of requests for access and a high hitting ratio is achieved. In the opposite direction, contents tend to spread the popularity index leading to gain low hitting ratio. With any scenario, P2C-only/P2C2 always give much better performance than TLRU/LFU.

c) *Impact of cache size on hitting ratio:* In Fig. 5a, NDN node uses P2C2 and the relative cache size is set in turn at 0.5%, 1.0%, 2.0% and 3.0%. The results show that hitting rate values at the final state are 45%, 55%, 63% and 63%



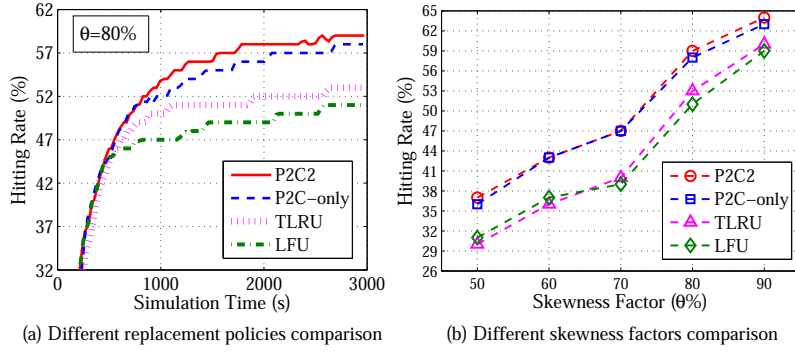


Fig. 4. Performance comparison with relative cache size 2%.

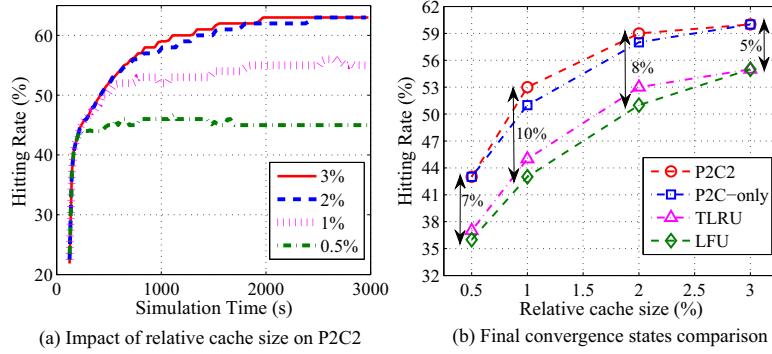


Fig. 5. Different relative cache sizes comparison.

respectively. From the above results, we notice that when we expand a cache volume, a higher hitting rate can be gotten. However, the growth of hitting rate does not linearly increase with cache volume. As shown in Fig. 5a, when cache size is large enough with relative cache size 2%, P2C2 can handle most of high ranking popular contents. If we continue to enlarge the cache volume to relative cache size 3%, the algorithm performance gain decreases.

Fig. 5b gives us a big picture about final coverage states of multiple replacement policies with multiple relative cache sizes. Again, P2C-only/P2C2 always outperform with higher hitting rate than TLRU/LFU in all situations. Moreover, with a simple cooperation among adjacent cells, P2C2 can gain more than 1% to 2% hitting value higher than P2C-only.

Fig. 5b also illustrates the impact of cache size on performance. The gap of hitting rate value between P2C2 and LFU is 10% at relative cache size 1%. When relative cache size increases to 2% and 3%, the slope of the curve turns to be flatter and the gap value reduces to 8% and 5%, respectively. Thus, there is a tradeoff between cache volume (cost) and algorithm performance, and we can make a balance to choose suitable cache size.

## V. CONCLUSION

We have presented a new type of cache decision and replacement policy applied for NDN. Our algorithm engages NDN nodes can achieve higher hitting rate, effective caching

and increase offloading server significantly. The simulation configuration has referred to the most recently existing papers and partitions that reflect the real world. We plan to further improve P2C2 through more complex and effective cooperation by sharing the whole multi-level prefix tree and dynamic cache size allocation. Then, in the near future, we will publish the NDN module to the OPNET community.

## REFERENCES

- [1] Cisco, "Cisco Annual Internet Report (2018 - 2023)", Feb. 2020.
- [2] Boubiche, Djallel Eddine, et al. "Mobile crowd sensing-Taxonomy, applications, challenges, and solutions". *Computers in Human Behavior* 101 (2019): 352-370.
- [3] Yao, Jingjing, Tao Han, and Nirwan Ansari. "On mobile edge caching". *IEEE Communications Surveys & Tutorials* 21.3 (2019): 2525-2553.
- [4] Appel, Gil, et al. "The future of social media in marketing". *Journal of the Academy of Marketing Science* 48.1 (2020): 79-95.
- [5] Ortíz, Jordi, Pedro Martínez-Julia, and Antonio Skarmeta. "Information-Centric Networking Future Internet Video Delivery". *User-Centric and Information-Centric Networking and Services: Access Networks, Storage and Cloud Perspective* (2019): 141.
- [6] Yu, Keping, et al. "Information-Centric Networking: Research and Standardization Status". *IEEE Access* 7 (2019): 126164-126176.
- [7] Scalable and Adaptive Internet Solutions (SAIL) project. Available: <http://www.sail-project.eu/>
- [8] Publish-Subscribe Internet Technology (PURSUIT) project. Available: <http://www.fp7-pursuit.eu/PursuitWeb/>
- [9] Named Data Networking. Available: <http://www.named-data.net/index.html>
- [10] Afanasyev, Alex, et al. "A brief introduction to Named Data Networking". *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*. IEEE, 2018.



- [11] Dehghan, Mostafa, et al. "A utility optimization approach to network cache design". *IEEE/ACM Transactions on Networking* 27.3 (2019): 1013-1027.
- [12] Zhang, Tiankui, et al. "Content-Centric Mobile Edge Caching". *IEEE Access* 8 (2019): 11722-11731.
- [13] Chen, Xing, et al. "Hit ratio driven mobile edge caching scheme for video on demand services". 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019.
- [14] Sun, Sanshan, et al. "Cooperative Caching with Content Popularity Prediction for Mobile Edge Caching". *Tehnicki vjesnik* 26.2 (2019): 503-509.
- [15] Malowidzki, Marek. "Network simulators: A developers perspective". *Proc. Int. Symp. Perform. Eval. Comput. Telecommun. Syst. SPECTS04*. 2004.
- [16] D. Rossi, G. Rossini, "Caching performance of content centric networks under multi-path routing (and more)", *Telecom ParisTech*, Technical report, Paris, France, 2011.
- [17] Li, Jun, et al. "Popularity-driven coordinated caching in named data networking". 2012 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS). IEEE, 2012.
- [18] OPNET Modeler. Available: [www.opnet.com](http://www.opnet.com)
- [19] Chen, Min, Yiming Miao, and Iztok Humar. "Introduction to OPNET Network Simulation". *OPNET IoT Simulation*. Springer, Singapore, 2019. 77-153.
- [20] NS-3 based Named Data Networking (NDN) simulator. Available: <http://ndnsim.net/index.html>

# Ritz Solution for Static Analysis of Thin-Walled Laminated Composite I-beams Based on First-Order Beam Theory

Ngoc-Duong Nguyen

Faculty of Civil Engineering

Ho Chi Minh City University of Technology and Education

Ho Chi Minh City, Vietnam

duongnn@hcmute.edu.vn

Trung-Kien Nguyen

Faculty of Civil Engineering

Ho Chi Minh City University of Technology and Education

Ho Chi Minh City, Vietnam

kiennt@hcmute.edu.vn

**Abstract**—Ritz solution is proposed for static analysis of thin-walled laminated composite I-beams in this paper. Beam model based on the first-order beam theory. The governing equations are obtained from Lagrange's equations. Ritz's approximation functions are developed to find the deflection of thin-walled beams under concentrated and uniform loads. Numerical results are presented and compared with those of available literature. Effects of fiber orientation, boundary conditions, and shear deformation on the displacement of thin-walled laminated composite I-beams are investigated. It can be seen that the Ritz method is efficient and straightforward for static analysis of thin-walled laminated composite I-beams

**Keywords**— Ritz method, Thin-walled composite beams, Static analysis, First-order beam theory.

## I. INTRODUCTION

Composite thin-walled structures are increasingly used in many engineering fields due to their high strength-to-weight and stiff-to-weight ratio. Thin-walled profile structures are produced in a variety of cross-sections, including channel, rectangular box, square tube, I-sections... in which I-section are becoming popular owing to convenience in production, connection, and erection.

The first thin-walled theory is proposed by Vlasov [1] for isotropic material. After that Bauld and Lih-Shyng [2] developed Vlasov's theory for thin-walled composite beams. Park et al. [3] predicted the deflection of thin-walled beams with an open section by Vlasov's theory. Lee and Lee [4] also analysed the static responses of thin-walled composite beams. It is seen that Vlasov's theory ignored shear strains, therefore, it is suitable for slender beams. For moderate beams, the shear effect becomes an important roll. Lee [5] developed a first-order beam theory (FOBT) for the bending response of thin-walled composite beams. In this study, a uniform load is considered for the deflection of beams. Back and Will [6] analysed bending and buckling behaviours of beams based on FOBT. Sheikh and Thomsen [7] developed a model beam considering shear effect to analyse the bending responses of beams. Kim et al. [8] applied FOBT to develop the static solution for thin-walled beams. It can be seen that considering shear deformation for bending analysis of thin-walled composite I-beams is still limited.

For the computational method, finite element method is used popularly for analysing thin-walled beams [9-11]. The dynamic stiff matrix is applied for the analysis of thin-walled

beams [12, 13]. Recently, Ritz method is developed to analyse free vibration and buckling responses of laminated composite I-beams [14]. Although, Ritz method is commonly used for beams with rectangular cross-section [15-18], it is interesting that Ritz method is rarely used to analyse bending behaviours of thin-walled beams.

The object of this paper is to apply Ritz method for bending analysis of thin-walled composite I-beams. The effects of fiber orientation, boundary conditions and anisotropic material on deflection of beams are investigated.

## II. THEORETICAL FORMULATION

### A. Kinematics

Consider a thin-walled beam with three coordinate systems as shown in Fig.1. Mid-surface displacement of beams ( $\bar{u}_1, \bar{u}_2, \bar{u}_3$ ) are defined as follows:

$$\bar{u}_1(s, z) = U_1(z) \sin \alpha(s) - U_2(z) \cos \alpha(s) - \phi(z) q(s) \quad (1)$$

$$\bar{u}_2(s, z) = U_1(z) \cos \alpha(s) + U_2(z) \sin \alpha(s) + \phi(z) r(s) \quad (2)$$

$$\bar{u}_3(s, z) = U_3(z) + \chi_x(z) x(s) + \chi_y(z) y(s) + \chi_\varpi(z) \varpi(s) \quad (3)$$

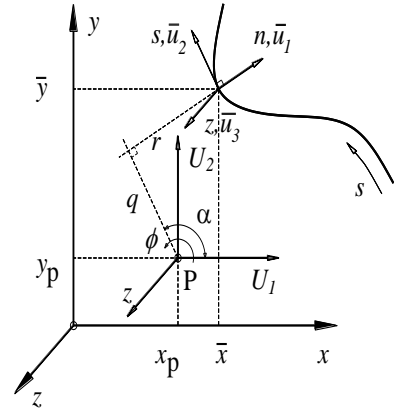


Fig. 1. Thin-walled coordinate systems

where  $U_1, U_2$  and  $U_3$  are displacements of pole point (P) in  $x$ -,  $y$ - and  $z$ - directions, respectively;  $\phi$  is the rotational angle

of cross-section about pole axis;  $\varpi(s) = \int_{s_0}^s r(s) ds$  is warping

function;  $\chi_x, \chi_y$  and  $\chi_\varpi$  are rotations of the cross-section with respect to  $x, y$  and  $\varpi$ , and defined by:

$$\chi_y = \gamma_{xz}^0 - U_1', \quad \chi_x = \gamma_{yz}^0 - U_2', \quad \chi_{\varpi} = \gamma_{\varpi}^0 - \phi' \quad (4)$$

where the prime superscript indicates differentiation with respect to  $z$ .  $\gamma_{xz}^0$ ,  $\gamma_{yz}^0$  and  $\gamma_{\varpi}^0$  are shear train and warping shear of thin-walled beams.

The displacements at any generic point  $(u_1, u_2, u_3)$  on the section of thin-walled beams are written by:

$$u_1(n, s, z) = \bar{u}_1(s, z) \quad (5)$$

$$u_2(n, s, z) = \bar{u}_2(s, z) + n\bar{\chi}_s(s, z) \quad (6)$$

$$u_3(n, s, z) = \bar{u}_3(s, z) + n\bar{\chi}_z(s, z) \quad (7)$$

where  $\bar{\chi}_s$  and  $\bar{\chi}_z$  are given by:

$$\bar{\chi}_z = \chi_y \sin \alpha - \chi_x \cos \alpha - \chi_{\varpi} q, \quad \bar{\chi}_s = -\frac{\partial \bar{u}_1}{\partial s} \quad (8)$$

The non-zero strains of beams are given by:

$$\varepsilon_z(n, s, z) = \bar{\varepsilon}_z(s, z) + n\bar{\kappa}_z(s, z) \quad (9)$$

$$= \varepsilon_z^0 + (x + n \sin \alpha) \kappa_y + (y - n \cos \alpha) \kappa_x + (\varpi - nq) \kappa_{\varpi}$$

$$\gamma_{sz}(n, s, z) = \bar{\gamma}_{sz}(s, z) + n\bar{\kappa}_{sz}(s, z) \quad (10)$$

$$\begin{aligned} &= \gamma_{sz}^0 \cos \alpha + \gamma_{yz}^0 \sin \alpha + \gamma_{\varpi}^0 r + n\kappa_{sz} \\ \gamma_{nz}(n, s, z) &= \bar{\gamma}_{nz}(s, z) + n\bar{\kappa}_{nz}(s, z) \quad (11) \\ &= \gamma_{nz}^0 \sin \alpha - \gamma_{yz}^0 \cos \alpha - \gamma_{\varpi}^0 q \end{aligned}$$

$$\text{where } \bar{\varepsilon}_z = \frac{\partial \bar{u}_3}{\partial z} = \varepsilon_z^0 + x\kappa_y + y\kappa_x + \varpi\kappa_{\varpi}, \quad (12)$$

$$\bar{\kappa}_z = \frac{\partial \bar{\chi}_z}{\partial z} = \kappa_y \sin \alpha - \kappa_x \cos \alpha - \kappa_{\varpi} q \quad (12)$$

$$\begin{aligned} \bar{\kappa}_{sz} &= \kappa_{sz}, \quad \bar{\kappa}_{nz} = 0, \quad \varepsilon_z^0 = U_3', \quad \kappa_x = \chi_x', \quad \kappa_y = \chi_y', \quad \kappa_{\varpi} = \chi_{\varpi}' \\ \kappa_{sz} &= \phi' - \chi_{\varpi} \end{aligned} \quad (13)$$

It can be seen that  $\varepsilon_z^0$ ,  $\kappa_x$ ,  $\kappa_y$ ,  $\kappa_{\varpi}$  and  $\kappa_{sz}$  are axial strain, biaxial curvatures in the  $x$ - and  $y$ - directions, warping curvature with respect to the shear center and twisting curvature in the beam, respectively.

### B. Constitutive equations

The stress and strain relations at the  $k^{\text{th}}$ -layer in  $(n, s, z)$  coordinate systems can be determined as:

$$\begin{Bmatrix} \sigma_z \\ \sigma_{sz} \\ \sigma_{nz} \end{Bmatrix}^{(k)} = \begin{pmatrix} \bar{Q}_{11}^* & \bar{Q}_{16}^* & 0 \\ \bar{Q}_{16}^* & \bar{Q}_{66}^* & 0 \\ 0 & 0 & \bar{Q}_{55}^* \end{pmatrix}^{(k)} \begin{Bmatrix} \varepsilon_z \\ \gamma_{sz} \\ \gamma_{nz} \end{Bmatrix} \quad (14)$$

$$\text{where: } \bar{Q}_{11}^* = \bar{Q}_{11} - \frac{\bar{Q}_{12}^2}{\bar{Q}_{22}}, \quad \bar{Q}_{16}^* = \bar{Q}_{16} - \frac{\bar{Q}_{12}\bar{Q}_{26}}{\bar{Q}_{22}},$$

$$\bar{Q}_{66}^* = \bar{Q}_{66} - \frac{\bar{Q}_{26}^2}{\bar{Q}_{22}}, \quad \bar{Q}_{55}^* = \bar{Q}_{55} \quad (15)$$

In Eq. (19),  $\bar{Q}_{ij}$  are the transformed reduced stiffnesses (refer [19] for more detail)

### C. Variational formulation

The strain energy  $\Pi_E$  of the system through volume  $\Omega$  is defined by:

$$\Pi_E = \frac{1}{2} \int_{\Omega} (\sigma_z \varepsilon_z + \sigma_{sz} \gamma_{sz} + \sigma_{nz} \gamma_{nz}) d\Omega \quad (16)$$

Substituting Eqs. (9), (10), (11) and (14) into Eq. (16) leads to:

$$\begin{aligned} \Pi_E &= \frac{1}{2} \int_0^L [E_{11} U_3'^2 + 2E_{16} U_3' U_1' + 2E_{17} U_3' U_2' + 2(E_{15} + E_{18}) U_3' \phi' \\ &+ 2E_{12} U_3' \chi_y' + 2E_{16} U_3' \chi_y' + 2E_{13} U_3' \chi_x' + 2E_{17} U_3' \chi_x + 2E_{14} U_3' \chi_{\varpi}' \\ &+ 2(E_{18} - E_{15}) U_3' \chi_{\varpi} + E_{66} U_1'^2 + 2E_{67} U_1' U_2' + 2(E_{56} + E_{68}) U_1' \phi' \\ &+ 2E_{26} U_1' \chi_y' + 2E_{66} U_1' \chi_y + 2E_{36} U_1' \chi_x' + 2E_{67} U_1' \chi_x + 2E_{46} U_1' \chi_{\varpi}' \\ &+ 2(E_{68} - E_{56}) U_1' \chi_{\varpi} + E_{77} U_2'^2 + 2(E_{57} + E_{78}) U_2' \phi' + 2E_{27} U_2' \chi_y' \\ &+ 2E_{67} U_2' \chi_y + 2E_{37} U_2' \chi_x' + 2E_{77} U_2' \chi_x + 2E_{47} U_2' \chi_{\varpi}' \\ &+ 2(E_{78} - E_{57}) U_2' \chi_{\varpi} + (E_{55} + 2E_{58} + E_{88}) \phi'^2 \\ &+ 2(E_{25} + E_{28}) \phi' \chi_y' + 2(E_{56} + E_{68}) \phi' \chi_y + 2(E_{35} + E_{38}) \phi' \chi_x' \\ &+ 2(E_{57} + E_{78}) \phi' \chi_x + 2(E_{45} + E_{48}) \phi' \chi_{\varpi}' + 2(E_{88} - E_{55}) \phi' \chi_{\varpi} \\ &+ E_{22} \chi_y'^2 + 2E_{26} \chi_y' \chi_y + E_{66} \chi_y^2 + 2E_{23} \chi_y' \chi_x' + 2E_{27} \chi_y' \chi_x \\ &+ 2E_{36} \chi_y' \chi_x + 2E_{67} \chi_y' \chi_x + 2E_{24} \chi_y' \chi_{\varpi}' + 2(E_{28} - E_{25}) \chi_y' \chi_{\varpi} \\ &+ 2E_{46} \chi_y' \chi_{\varpi} + 2(E_{68} - E_{56}) \chi_y \chi_{\varpi} + E_{33} \chi_x'^2 + 2E_{37} \chi_x' \chi_x \\ &+ E_{77} \chi_x^2 + 2E_{34} \chi_x' \chi_{\varpi}' + 2(E_{38} - E_{35}) \chi_x \chi_{\varpi} + 2E_{47} \chi_x \chi_{\varpi}' \\ &+ 2(E_{78} - E_{57}) \chi_x \chi_{\varpi} + E_{44} \chi_{\varpi}'^2 + 2(E_{48} - E_{45}) \chi_{\varpi}' \chi_{\varpi} \\ &+ (E_{88} - 2E_{58} + E_{55}) \chi_{\varpi}^2] dz \end{aligned} \quad (17)$$

where  $L$  and  $E_{ij}$  are length and stiffness coefficients of thin-walled composite beams. The work done  $\Pi_W$  of the system by uniform load  $q_y$  and concentrated load  $P_y$  applied at  $z_L$  can be expressed:

$$\Pi_W = \int_0^L q_y V dz + P_y V(z_L) \quad (18)$$

The total potential energy of the system is obtained by:

$$\Pi = \Pi_E - \Pi_W \quad (19)$$

### D. Ritz method

The displacement fields of the thin-walled composite beams are approximated by using Ritz's approximation functions:

$$\begin{aligned} U_1(z) &= \sum_{j=1}^m \varphi_j(z) U_{1j}, \quad U_2(z) = \sum_{j=1}^m \varphi_j(z) U_{2j}, \\ U_3(z) &= \sum_{j=1}^m \varphi_j'(z) U_{3j}, \quad \phi(z) = \sum_{j=1}^m \varphi_j(z) \phi_j, \quad \chi_y(z) = \sum_{j=1}^m \varphi_j'(z) \chi_{yj}, \\ \chi_x(z) &= \sum_{j=1}^m \varphi_j'(z) \chi_{xj}, \quad \chi_{\varpi}(z) = \sum_{j=1}^m \varphi_j'(z) \chi_{\varpi j} \end{aligned} \quad (20)$$

where  $U_{1j}, U_{2j}, U_{3j}, \phi_j, \chi_{xj}, \chi_{yj}$  and  $\chi_{\varpi j}$  are Ritz's parameters, which need to be determined and  $\varphi_j(z)$  are Ritz's approximation functions, as seen in Table 1. In this paper, four typical boundary conditions (BCs) as clamped-

clamped (C-C), clamped-simply supported (C-S) simply-supported (S-S) and clamped-free (C-F) are considered.

TABLE I. RITZ'S FUNCTIONS FOR VARIOUS BOUNDARY CONDITIONS

B C	$\frac{\varphi_j(z)}{e^{\frac{-z}{L}}}$	$z=0$	$z=L$
S- S	$\frac{z}{L}\left(1-\frac{z}{L}\right)$	$U_1=U_2=\phi=0$	$U_1=U_2=\phi=0$
C- F	$\left(\frac{z}{L}\right)^2$	$U_1=U_2=\phi=0$ $U_1'=U_2'=\phi'=0$ $U_3=\chi_y=\chi_x=\chi_\sigma=0$	
C- S	$\left(\frac{z}{L}\right)^2\left(1-\frac{z}{L}\right)$	$U_1=U_2=\phi=0$ $U_1'=U_2'=\phi'=0$ $U_3=\chi_y=\chi_x=\chi_\sigma=0$	$U_1=U_2=\phi=0$
C- C	$\left(\frac{z}{L}\right)^2\left(1-\frac{z}{L}\right)^2$	$U_1=U_2=\phi=0$ $U_1'=U_2'=\phi'=0$ $U_3=\chi_y=\chi_x=\chi_\sigma=0$	$U_1=U_2=\phi=0$ $U_1'=U_2'=\phi'=0$ $U_3=\chi_y=\chi_x=\chi_\sigma=0$

By substituting Eq. (20) into Eq. (19), Lagrange's equations are used to formulate the governing equations:

$$\frac{\partial \Pi}{\partial p_j} = 0 \quad (21)$$

with  $p_j$  representing the values of  $(U_{1j}, U_{2j}, U_{3j}, \phi_j, \chi_{xj}, \chi_{yj}, \chi_{\sigma j})$ , the bending response of the beams can be determined by solving the following equation:

$$\begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} & \mathbf{K}^{13} & \mathbf{K}^{14} & \mathbf{K}^{15} & \mathbf{K}^{16} & \mathbf{K}^{17} \\ {}^T\mathbf{K}^{12} & \mathbf{K}^{22} & \mathbf{K}^{23} & \mathbf{K}^{24} & \mathbf{K}^{25} & \mathbf{K}^{26} & \mathbf{K}^{27} \\ {}^T\mathbf{K}^{13} & {}^T\mathbf{K}^{23} & \mathbf{K}^{33} & \mathbf{K}^{34} & \mathbf{K}^{35} & \mathbf{K}^{36} & \mathbf{K}^{37} \\ {}^T\mathbf{K}^{14} & {}^T\mathbf{K}^{24} & {}^T\mathbf{K}^{34} & \mathbf{K}^{44} & \mathbf{K}^{45} & \mathbf{K}^{46} & \mathbf{K}^{47} \\ {}^T\mathbf{K}^{15} & {}^T\mathbf{K}^{25} & {}^T\mathbf{K}^{35} & {}^T\mathbf{K}^{45} & \mathbf{K}^{55} & \mathbf{K}^{56} & \mathbf{K}^{57} \\ {}^T\mathbf{K}^{16} & {}^T\mathbf{K}^{26} & {}^T\mathbf{K}^{36} & {}^T\mathbf{K}^{46} & {}^T\mathbf{K}^{56} & \mathbf{K}^{66} & \mathbf{K}^{67} \\ {}^T\mathbf{K}^{17} & {}^T\mathbf{K}^{27} & {}^T\mathbf{K}^{37} & {}^T\mathbf{K}^{47} & {}^T\mathbf{K}^{57} & {}^T\mathbf{K}^{67} & \mathbf{K}^{77} \end{bmatrix} \begin{bmatrix} \mathbf{U}_3 \\ \mathbf{U}_1 \\ \mathbf{U}_2 \\ \Phi \\ \chi_y \\ \chi_x \\ \chi_\sigma \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{F} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (22)$$

The explicit forms of stiffness matrix  $\mathbf{K}$  and force vector  $\mathbf{F}$  are given as followings:

$$K_{ij}^{11} = E_{11} \int_0^L \varphi_i'' \varphi_j'' dz, \quad K_{ij}^{12} = E_{16} \int_0^L \varphi_i'' \varphi_j' dz, \quad K_{ij}^{13} = E_{17} \int_0^L \varphi_i'' \varphi_j' dz,$$

$$K_{ij}^{14} = (E_{15} + E_{18}) \int_0^L \varphi_i'' \varphi_j' dz, \quad K_{ij}^{15} = E_{12} \int_0^L \varphi_i'' \varphi_j' dz + E_{16} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{16} = E_{13} \int_0^L \varphi_i'' \varphi_j'' dz + E_{17} \int_0^L \varphi_i'' \varphi_j' dz,$$

$$K_{ij}^{17} = E_{14} \int_0^L \varphi_i'' \varphi_j'' dz + (E_{18} - E_{15}) \int_0^L \varphi_i'' \varphi_j' dz, \quad K_{ij}^{22} = E_{66} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{23} = E_{67} \int_0^L \varphi_i' \varphi_j' dz, \quad K_{ij}^{24} = (E_{56} + E_{68}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{25} = E_{26} \int_0^L \varphi_i' \varphi_j'' dz + E_{66} \int_0^L \varphi_i' \varphi_j' dz$$

$$K_{ij}^{26} = E_{36} \int_0^L \varphi_i' \varphi_j'' dz + E_{67} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{27} = E_{46} \int_0^L \varphi_i' \varphi_j'' dz + (E_{68} - E_{56}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{33} = E_{77} \int_0^L \varphi_i' \varphi_j' dz + N_0 \int_0^L \varphi_i' \varphi_j' dz, \quad K_{ij}^{34} = (E_{57} + E_{78}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{35} = E_{27} \int_0^L \varphi_i' \varphi_j'' dz + E_{67} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{36} = E_{37} \int_0^L \varphi_i' \varphi_j'' dz + E_{77} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{37} = E_{47} \int_0^L \varphi_i' \varphi_j'' dz + (E_{78} - E_{57}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{44} = (E_{55} + 2E_{58} + E_{88}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{45} = (E_{25} + E_{28}) \int_0^L \varphi_i' \varphi_j'' dz + (E_{56} + E_{68}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{46} = (E_{35} + E_{38}) \int_0^L \varphi_i' \varphi_j'' dz + (E_{57} + E_{78}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{47} = (E_{45} + E_{48}) \int_0^L \varphi_i' \varphi_j'' dz + (E_{88} - E_{55}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{55} = E_{22} \int_0^L \varphi_i'' \varphi_j'' dz + E_{26} \int_0^L (\varphi_i'' \varphi_j' + \varphi_i' \varphi_j'') dz + E_{66} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{56} = E_{23} \int_0^L \varphi_i'' \varphi_j'' dz + E_{27} \int_0^L \varphi_i'' \varphi_j' dz + E_{36} \int_0^L \varphi_i' \varphi_j'' dz + E_{67} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{57} = E_{24} \int_0^L \varphi_i'' \varphi_j'' dz + (E_{28} - E_{25}) \int_0^L \varphi_i'' \varphi_j' dz + E_{46} \int_0^L \varphi_i' \varphi_j'' dz$$

$$+ (E_{68} - E_{56}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{66} = E_{33} \int_0^L \varphi_i'' \varphi_j'' dz + E_{37} \int_0^L (\varphi_i'' \varphi_j' + \varphi_i' \varphi_j'') dz + E_{77} \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{67} = E_{34} \int_0^L \varphi_i'' \varphi_j'' dz + (E_{38} - E_{35}) \int_0^L \varphi_i'' \varphi_j' dz + E_{47} \int_0^L \varphi_i' \varphi_j'' dz$$

$$+ (E_{78} - E_{57}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$K_{ij}^{77} = E_{44} \int_0^L \varphi_i'' \varphi_j'' dz + (E_{48} - E_{45}) \int_0^L (\varphi_i'' \varphi_j' + \varphi_i' \varphi_j'') dz$$

$$+ (E_{88} - 2E_{58} + E_{55}) \int_0^L \varphi_i' \varphi_j' dz,$$

$$F_i = \int_0^L q_y \varphi_j dz + P_y \varphi_j(z_L) \quad (23)$$

### III. NUMERICAL RESULTS

This section is outlined as follows: firstly, the convergence study is carried out to test Ritz's approximation functions. Secondly, verification is examined to evaluate the effectiveness and accuracy of the present solution. Finally, the thin-walled composite I-beams are considered to investigate the effects of fiber angle, BCs, and anisotropic material on the deflection of beams. Material properties of beams are assumed as:  $E_1 = 53780 \text{ MPa}$ ,  $E_2 = E_3 = 17930 \text{ MPa}$ ,

$G_{23} = 3450 \text{ MPa}$  ,  $G_{12} = G_{13} = 8960 \text{ MPa}$  ,  $\nu_{23} = 0.34$  ,  $\nu_{12} = \nu_{13} = 0.25$  . Cross-section of I-beams are shown in Fig. 2 with  $b_1 = b_2 = b_3 = 0.05 \text{ m}$  and  $h_1 = h_2 = h_3 = 0.00208 \text{ m}$  .

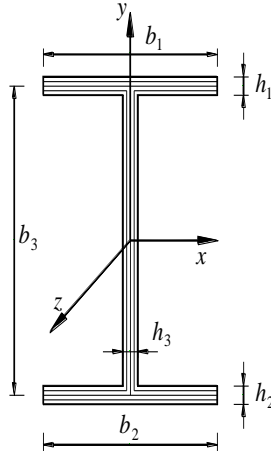


Fig. 2. Geometry of I-beam

#### A. Convergence study

To test the convergence of the proposed solution, the I-beams ( $L = 2.5 \text{ m}$ ) subjected to a uniformly distributed load ( $p_y = 1 \text{ kN/m}$ ) are considered. The web and flanges of I-beams are assumed to be symmetrically laminated angle-ply  $[60/-60]_{4s}$  with respect to its mid-plane. The mid-span deflections of I-beams with various BCs are displayed in Table 2. It is seen that the proposed solution converges at  $m = 10$ , and this number is selected for the following examples. In comparison, the present functions converge more quickly than trigonometric solution [15].

TABLE II. MID-SPAN DEFLECTION OF COMPOSITE I-BEAMS SUBJECTED TO A UNIFORM LOAD (cm).

BC	$m$					
	2	4	6	8	10	12
S-S	16.392	16.896	17.005	16.997	16.998	16.998
C-F	55.907	57.629	57.783	57.775	57.777	57.777
C-S	6.405	6.845	6.818	6.820	6.820	6.821
C-C	3.437	3.424	3.426	3.427	3.427	3.428

#### B. Example 1

For the purpose verification of the present solution, the simply-supported ( $L = 2.5 \text{ m}$ ) and cantilever ( $L = 1 \text{ m}$ ) I-beams with configuration as the previous example are considered. Deflection at mid-span of simply-supported beams subjected to uniform load ( $p_y = 1 \text{ kN/m}$ ) are displayed in Table 3, and deflection at free end of cantilever beams subjected to concentrated load ( $P_y = 1 \text{ kN}$ ) are displayed in Table 4. It can be observed from these figures that the result presents are agreement with Lee [5], Sheikh and Thomsen [7], Back and Will [6], Kim et al. [8], which based on first-order beam theory and solved by FEM, and slightly larger than results of Lee and Lee [4], Park et al. [3] because of considering shear effect.

TABLE III. MID-SPAN DEFLECTION OF SIMPLY-SUPPORTED COMPOSITE I-BEAMS SUBJECTED TO A UNIFORM LOAD (cm).

Lay-up	Reference				
	Present (Shear)	Lee [5] (Shear)	Lee and Lee [4] (No shear)	Sheikh and Thomse n [7] (Shear)	Back and Will [6] (Shear)
$[0]_{16}$	6.261	6.259	6.233	6.264	6.261
$[15/-15]_{4s}$	6.926	6.923	6.899	6.929	6.926
$[30/-30]_{4s}$	9.317	9.314	9.290	9.320	9.317
$[45/-45]_{4s}$	13.450	13.446	13.421	13.450	13.450
$[60/-60]_{4s}$	16.998	16.992	16.962	17.000	17.000
$[75/-75]_{4s}$	18.455	18.449	18.411	18.460	18.460
$[0/90]_{4s}$	9.383	9.381	9.299	9.387	9.384

TABLE IV. DEFLECTION AT FREE END OF CANTILEVER COMPOSITE I-BEAMS SUBJECTED TO CONCENTRATED LOAD (cm).

Lay-up	Reference			
	Present (Shear)	Kim et al. [8] (Shear)	Kim et al. [8] (No shear)	Park et al. [3] (No shear)
$[15/-15]_{4s}$	4.555	4.611	4.519	4.521
$[30/-30]_{4s}$	6.122	6.156	6.084	6.089
$[45/-45]_{4s}$	8.833	8.855	8.785	8.795
$[60/-60]_{4s}$	11.162	11.18	11.10	11.12
$[75/-75]_{4s}$	12.122	12.16	12.05	12.07
$[0/90]_{4s}$	6.171	6.201	6.089	6.093

Mid-span deflection of beams subjected to uniformly distributed load ( $p_y = 1 \text{ kN/m}$ ) with various ply-angle are shown in Table 5. It is seen from this table that the deflections increase as ply-angles increase, and is the largest for beam with C-F BC and the smallest for one with C-C BC.

TABLE V. MID-SPAN DEFLECTION OF COMPOSITE I-BEAMS SUBJECTED TO A UNIFORM LOAD (cm).

Lay-up	BC		
	C-F	C-S	C-C
$[0]_{16}$	21.274	2.521	1.274
$[15/-15]_{4s}$	23.535	2.786	1.406
$[30/-30]_{4s}$	31.666	3.742	1.884
$[45/-45]_{4s}$	45.718	5.398	2.713
$[60/-60]_{4s}$	57.777	6.820	3.427
$[75/-75]_{4s}$	62.728	7.408	3.724
$[0/90]_{4s}$	31.889	3.774	1.904

### C. Example 2

This example investigates the effect of anisotropic material on the deflection of thin-walled beams. Material properties of beams are assumed  $E_1/E_2 = \text{open}$ ,  $E_2 = E_3$ ,  $G_{23} = 0.2E_2$ ,  $G_{12} = G_{13} = 0.5E_2$ ,  $\nu_{12} = \nu_{13} = \nu_{23} = 0.25$ . For convenience, the non-dimensional deflection is used as follows:

$$\bar{V} = \frac{VE_2 b_3^3}{q_y L^4}. \text{ Fig. 3a and b display variation of non-}$$

dimensional deflections at mid-span of beams with respect to  $E_1/E_2$  ratio for beams with  $[15/-15]_{4S}$  and  $[75/-75]_{4S}$ , respectively. It is seen from Fig 3 that for beam with  $[15/-15]_{4S}$ , deflections decrease as  $E_1/E_2$  ratio increases, however, for beam with  $[75/-75]_{4S}$  variation of  $E_1/E_2$  ratio hardly effects to deflections.

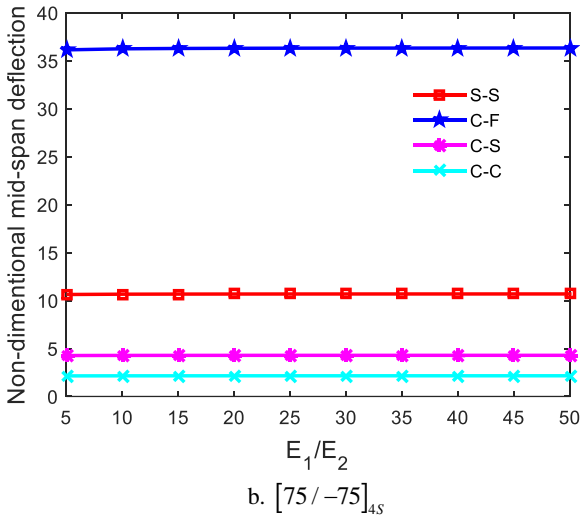
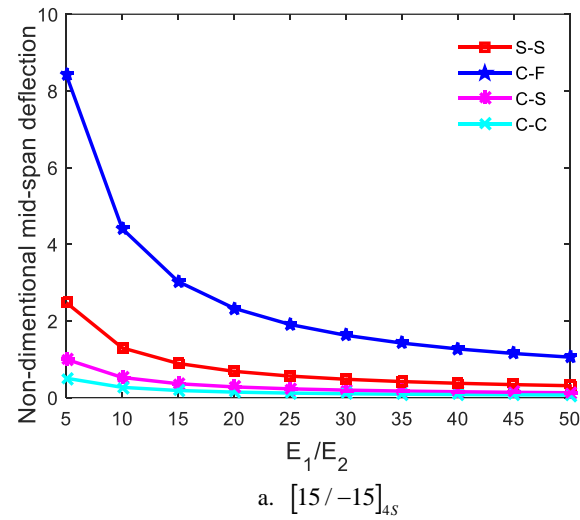


Fig. 3. Variation of non-dimensional deflection of beams with respect to  $E_1/E_2$  ratio

### IV. CONCLUSION

Bending behaviours of thin-walled composite I-beams are investigated in this paper. Displacement fields of beams based on FOBT. Lagrange's equation is used to obtain governing equations. Ritz method is developing to solve problem.

Numerical results are compared with those of available literature. It can be concluded that Ritz method is simple and effective for bending responses of thin-walled composite I-beams.

### REFERENCES

- [1] V. Vlasov, Thin-walled elastic beams. Israel program for scientific translations, Jerusalem. 1961, Oldbourne Press, London.
- [2] N.R. Bauld and T. Lih-Shyng, A Vlasov theory for fiber-reinforced beams with thin-walled open cross sections, International Journal of Solids and Structures. 20(3) (1984) 277-297.
- [3] Y. Park, H. Kwon, and D. Shin, Bending analysis of symmetrically laminated composite open section beam by Vlasov-type thin-walled beam theory, Korean Society of Civil Engineers Journal. 20(1) (2000) 125-141.
- [4] J. Lee and S.-h. Lee, Flexural-torsional behavior of thin-walled composite beams, Thin-Walled Structures. 42(9) (2004) 1293-1305.
- [5] J. Lee, Flexural analysis of thin-walled composite beams using shear-deformable beam theory, Composite Structures. 70(2) (2005) 212-222.
- [6] S.Y. Back and K.M. Will, Shear-flexible thin-walled element for composite I-beams, Engineering Structures. 30(5) (2008) 1447-1458.
- [7] A.H. Sheikh and O.T. Thomsen, An efficient beam element for the analysis of laminated composite beams of thin-walled open and closed cross sections, Composites Science and Technology. 68(10) (2008) 2273-2281.
- [8] N.-I. Kim, C.-K. Jeon, and J. Lee, A new laminated composite beam element based on eigenvalue problem, European Journal of Mechanics-A/Solids. 41 (2013) 111-122.
- [9] N.-I. Kim and J. Lee, Nonlinear analysis of thin-walled Al/Al<sub>2</sub>O<sub>3</sub> FG sandwich I-beams with mono-symmetric cross-section, European Journal of Mechanics-A/Solids. 69 (2018) 55-70.
- [10] V. Niki, Shear-deformable hybrid finite element method for buckling analysis of composite thin-walled members. 2018.
- [11] T.P. Vo and J. Lee, Geometrical nonlinear analysis of thin-walled composite beams using finite element method based on first order shear deformation theory, Archive of Applied Mechanics. 81(4) (2011) 419-435.
- [12] N.-I. Kim, Shear deformable composite beams with channel-section on elastic foundation, European Journal of Mechanics-A/Solids. 36 (2012) 104-121.
- [13] N.-I. Kim, Shear deformable doubly-and mono-symmetric composite I-beams, International Journal of mechanical sciences. 53(1) (2011) 31-41.
- [14] N.-D. Nguyen, T.-K. Nguyen, T.P. Vo, T.-N. Nguyen, and S. Lee, Vibration and buckling behaviours of thin-walled composite and functionally graded sandwich I-beams, Composites Part B: Engineering. 166 (2019) 414-427.
- [15] T.-K. Nguyen, N.-D. Nguyen, T.P. Vo, and H.-T. Thai, Trigonometric-series solution for analysis of laminated composite beams, Composite Structures. 160 (2017) 142-151.
- [16] N.-D. Nguyen, T.-K. Nguyen, T.-N. Nguyen, and H.-T. Thai, New Ritz-solution shape functions for analysis of thermo-mechanical buckling and vibration of laminated composite beams, Composite Structures. 184 (2017) 452-460.
- [17] J. Mantari and F. Canales, Free vibration and buckling of laminated beams via hybrid Ritz solution for various penalized boundary conditions, Composite Structures. 152 (2016) 306-315.
- [18] M. Şimşek, Static analysis of a functionally graded beam under a uniformly distributed load by Ritz method, International Journal of Engineering and Applied Sciences. 1(3) (2009) 1-11.
- [19] J.N. Reddy, Mechanics of laminated composite plates and shells: theory and analysis (CRC press, 2004).



# A Redundant Unit Form of Quasi-Z-source T-Type Inverter with Fault-Tolerant Capability

Duc-Tri Do

*Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Viet Nam  
tridd@hcmute.edu.vn*

Vinh-Thanh Tran

*Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Viet Nam  
tranvinhthanh.tc@gmail.com*

Hieu-Giang Le

*Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Viet Nam  
gianglh@hcmute.edu.vn*

Thanh-Hai Quach

*Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Viet Nam  
haiqt@hcmute.edu.vn*

Viet-Anh Truong

*Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Viet Nam  
anhvt@hcmute.edu.vn*

Minh-Khai Nguyen, *Senior Member,  
IEEE*

*Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Viet Nam  
nmkhai00@gmail.com*

Thi-Ngoc-Han Vuong

*Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam  
hanvtn@hcmute.edu.vn*

**Abstract**— A new combination between a three-phase three-level quasi-Z-source inverter and redundant unit form which ensured the operation of the T-type inverter in open-circuit fault is introduced in this paper. By consisting of the quasi-Z source network, this topology has some advantages such as: behaving as a buck-boost converter and shoot-through immunity. Furthermore, the redundant unit added into traditional T-type inverter helps to solve the open-circuit fault occurring in semiconductor devices of the converter without degrading the peak-value of the load current or increasing the DC-link voltage like traditional open-circuit fault-tolerant scheme. The PWM technique and the operating principle are detailed in this paper. The effectiveness of the inverter is verified through simulation results by using MATLAB Simulink.

**Keywords**— *Quasi-Z source, fault-tolerance, reliability evaluation, T-Type inverter, three-level inverter, buck-boost inverter, redundant unit form.*

## I. INTRODUCTION

Because of the advantages of Z-source (ZS) inverter (ZSI) compared to traditional voltage source inverter (VSI), this ZSI is applied to many industrial applications such as: photovoltaic (PV) grid-connected systems, motor drive, etc [1], [2]. The ZSI was explored first by Prof. F. Z. Peng in 2003, which uses one diode, two inductors and two capacitors to form the ZS network [3]. In this work, the two-level inverter was considered to install after the ZS structure to convert DC input voltage to three-phase AC voltage. By adopting impedance-source network, this topology behaves as a buck-boost converter without using additional DC-DC converter or AC-AC transformer. In addition, this structure can solve the short-through (ST) problem in conventional VSI without using dead-time which lead to reduce the efficiency of the inverter. Inheriting the superior of ZSI, the literature in [4] presented a new pulse width modulation (PWM) strategy to maximum the voltage gain of ZSI. In this

study, two PWM schemes were proposed, one is simple sine PWM technique and the other is third harmonic injection technique. In this work, all the time interval of zero vector is replaced by ST vector which used to boost the DC input voltage. The PWM strategy minimizing the inductor current ripple was proposed in [5]. In this literature, the ST state is divided into several equal parts, thus there are more than two time periods that the ZS network is stored energy in one switching period. As a result, the inductor current ripple is significantly decreased. However, these studies were based on two-level inverter, so the quality of output voltage as well as output current are not good compared to multilevel inverter. The literature in [6] presented a new topology which connects two identical ZS networks in series in order to create three output terminal of intermediate network which feeds to the three phase three level neutral point clamped (NPC) inverter to improve the output quality. Moreover, in this study, the common-mode voltage reduction was discussed. In this design, a large number of passive components were utilized, thus the cost and volume of the system are increased, consequently. To improve the limitation of [6], a new topology using only one ZS network incorporated with NPC inverter was presented in [7]. This topology required one split DC source feeding intermediate network. The mid-point of input power supply and the output of the ZS network provide a three-level voltage at output terminal of the inverter. As a result, the quality of output waveform is significantly improved.

However, the literature in [8] figures out that this topology of intermediate network has some drawbacks such as: high stress on power devices, discontinuous input current. For that reason, the quasi Z-source (qZS) inverter (qZSI) was introduced to overcome these limitations [9]. Because of the advantages of qZSI compared to ZSI, this topology has been attracted a plenty of researchers in the world. In [10], a

combination between qZS network and three-level T-type inverter (3L-T<sup>2</sup>I) was introduced to provide some advantages like: high output quality, low count of passive components at inverter branch. Moreover, the space vector PWM technique was proposed for this configuration to reduce common-mode voltage (CMV). Because of applying the qZS network for 3L-T<sup>2</sup>I, this combination is suitable for low and medium voltage application. However, the stress on capacitor of impedance network is still high. The literature in [11] incorporates the second type of qZS network to reduce the stress on capacitors whereas the output quality is still maintained.

With the development of industry, the reliability of the system has been become as one of the most important issues of the inverter. The literature [12] figured out that the main reason causing system failure is semiconductor failure. This failure can be classified into two case: open-circuit (OC) failure and short-circuit (SC) failure. The SC failure is more serious than the OC failure. However, the SC fault can be changed to the OC fault by using high speed fuse [13]. In [13], the qZS network was used with the 3L-T<sup>2</sup>I for fault-tolerant capability. The PWM strategy for fault-tolerant scheme (high-side and low-side switch OC fault as well as bi-directional switch OC fault) were discussed. This PWM technique changed the modulation index and ST duty ratio to guarantee the output capacity when high-side and low-side switch OC faults occur. As a result, the stress on power devices is significantly increased in fault condition. To improve the stress on power devices as well as output quality, a new topology which used additional switching devices added to traditional 3L-T<sup>2</sup>I to form redundant unit was proposed in [14]. However, this configuration behaves as a buck converter so the voltage gain is not high enough for many applications which requires high voltage rate at the output.

In this paper, a new combination between the fault tolerant 3L-T<sup>2</sup>I introduced in [14] with qZS network was proposed. This topology ensures that the converter can operate under OC semiconductor failure condition without increasing voltage stress on power devices. Moreover, this configuration can also ensure the buck-boost capability and other advantages of qZSI. The operating principle of this topology is presented and verified by MATLAB Simulink.

## II. THE qZST<sup>2</sup>I UNDER SEMICONDUCTOR FAILURE MODES

The configuration of the proposed qZST<sup>2</sup>I is shown in Fig. 1. In this structure, an intermediate network, which consists of two identical qZS networks, is placed before the inverter leg to boost the DC-link voltage. This network consists of three inductors, four capacitors and two diodes, as presented in Fig. 1. The inverter branch is formed by combining the 3L-T<sup>2</sup>I with a redundant leg which consists of two switches and six diodes. Different from conventional 3L-T<sup>2</sup>I which uses two switches connected in series to form the bi-directional switch, this configuration uses two diodes and two active switches to generate the bi-directional switch, as illustrated in Fig. 1, which is easy for connecting the redundant branch to provide the fault-tolerant capability.

The same as other single-state converter, this topology has two main mode during operation which are ST mode and non-ST (NST) mode. The ST mode is generated by triggered "ON" all switches of the inverter leg at the same time. As a result, the qZS network stores energy from the input DC supply

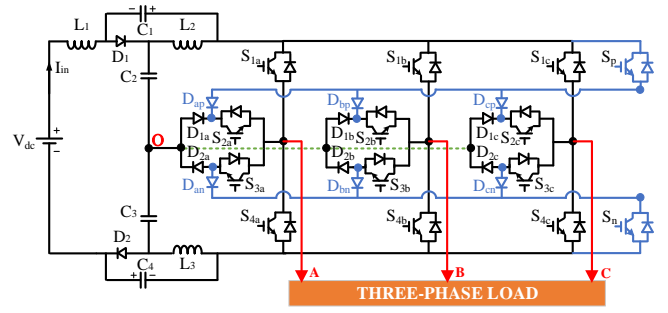


Fig. 1. Three-level qZST<sup>2</sup>I.

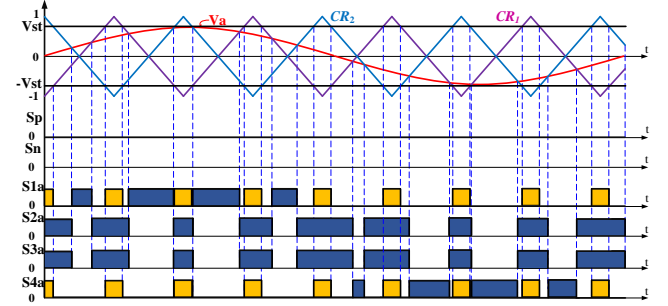


Fig. 2. PWM modulation strategy for phase A of the proposed inverter in normal condition.

whereas the output voltage is zero. Therefore, in order not to generate the distortion at output voltage, this state is inserted into zero state of the inverter leg. Because of using this state, this type of inverter configuration does not need the dead-time to limit the effect of ST phenomenon like conventional VSI. During NST mode, the qZS network provides energy for the inverter branch, the inverter leg operates like traditional three-level inverter by producing three-level voltage at the output terminal which are  $+V_{DC-link}/2$ , zero and  $-V_{DC-link}/2$  where the  $V_{DC-link}$  is the output voltage of intermediate network, as illustrated in Table I.

TABLE I. SWITCHING STATES OF PROPOSED TOPOLOGY (X = A, B, C)

Mode	"ON" switches	"ON" diode	$V_{xo}$
ST	$S_{1x}, S_{2x}, S_{3x}, S_{4x}$	$D_{1x}, D_{2x}$	0
NST	$S_{1x}$	$D_{1x}, D_{2x}$	$+V_{DC-link}/2$
	$S_{2x}, S_{3x}$	$D_{1x}, D_{2x}$	0
	$S_{4x}$	$D_{1x}, D_{2x}$	$-V_{DC-link}/2$

The same as other qZSI, the DC-link voltage of this topology can be expressed as [10]:

$$V_{DC-link} = \frac{V_{dc}}{1 - 2D_0} \quad (1)$$

Where  $V_{DC-link}$ : the output voltage of qZS network  
 $V_{dc}$ : the input DC supply  
 $D_0$ : the ST duty ratio.

The RMS value of output load voltage can be calculated as:

$$V_{x,RMS} = \frac{1}{2\sqrt{2}} m V_{DC-link} = \frac{1}{2\sqrt{2}} m \frac{V_{dc}}{1 - 2D_0} \quad (2)$$

Where  $V_{x,RMS}$ : the RMS value of output load voltage  
 $m$ : modulation index

In order not to affect the output voltage the relationship between the modulation index and ST duty ratio must be:

$$\begin{cases} m \leq 1 \\ m + D_0 \leq 1 \end{cases} \quad (3)$$

The control signal of the switches of phase A is generated by using two high frequency carriers which are  $CR_1$  and  $CR_2$ ,

and one sine reference signal  $V_a$ , as shown in Fig. 2. The reference signals of three-phase are expressed as:

$$\begin{cases} V_a = m \cdot \sin(\theta) \\ V_b = m \cdot \sin(\theta - 2\pi/3) \\ V_c = m \cdot \sin(\theta + 2\pi/3) \end{cases} \quad (4)$$

Due to the symmetry of the 3L-T<sup>2</sup>I the OC faults occurring at high-side switches  $S_{1x}$  can be handled by the same way compared to the low-side switches  $S_{4x}$ . Therefore, this paper just focuses on the fault-tolerant solution for high-side switches and bi-directional switches. Generally, assume that the OC fault is appeared at  $S_{1a}$  or bi-directional switch ( $S_{2a}$  or/and  $S_{3a}$ ).

#### Case 1: Open-circuit fault in $S_{1a}$

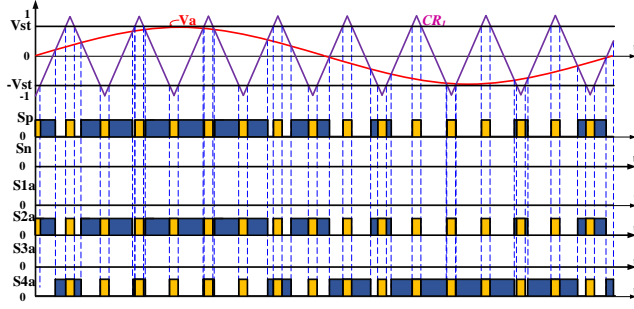


Fig. 3. PWM modulation strategy in post-fault condition of  $S_{1a}$

When the OC fault occurring at  $S_{1a}$ , the inverter leg cannot generate the  $+V_{DC-link}/2$  at output terminal of phase A. Therefore, the output voltage of phase A is zero during positive half period which significantly affects the quality of output load voltage as well as current. To solve this problem, the introduced configuration uses the switch  $S_p$ , diode  $D_{ap}$  and switch  $S_{2a}$  to replace the switch  $S_{1a}$ . Thus, instead of triggering the  $S_{1a}$ , the switch  $S_{ap}$  and  $S_{2a}$  are triggered “ON” at the same time, as a result, diode  $D_{ap}$  is forward bias and the value  $+V_{PN}/2$  is generated at output terminal. The PWM strategy of the inverter is illustrated in Fig. 3. The other healthy phases use high switch and low switch to ensures the two-level inverter operation.

#### Case 2: Open-circuit fault in $S_{2a}$ and/or $S_{3a}$

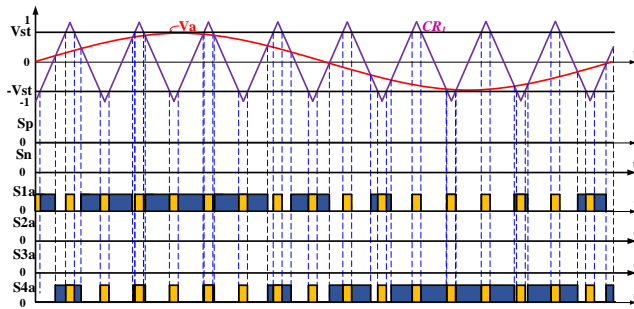


Fig. 4. PWM modulation strategy in post-fault condition of  $S_{2a}$

In this case of OC fault, the phase A cannot achieve the value zero at its output terminal. This problem causes the distortion at output waveform. It can be solved by applying the modified PWM method illustrated in Fig. 4. In this condition, the faulty-phase behaves as a two-level inverter by using one reference signal ( $V_a$ ) and one carrier ( $CR_1$ ) to generate the control signal of switches of phase A. The other healthy phases maintain the operation like normal condition.

TABLE II. PARAMETER USED IN SIMULATION

Parameter/Component		Attributes
Input voltage	$V_{dc}$	180 V
Output frequency	$f_o$	50 Hz
Carrier frequency	$f_s$	5 kHz
ST duty ratio	$D_0$	0.3
Modulation index	$M$	0.7
Desired output load voltage	$V_{o,RMS}$	110 V <sub>RMS</sub>
Boost inductor	$L_1 = L_2 = L_3$	3mH
Capacitors	$C_1 = C_2 = C_3 = C_4$	1000 $\mu$ F
Three-phase low-pass filter	$L_f$ and $C_f$	3mH and 10 $\mu$ F
Three-phase resistor load	$R_{load}$	40 $\Omega$

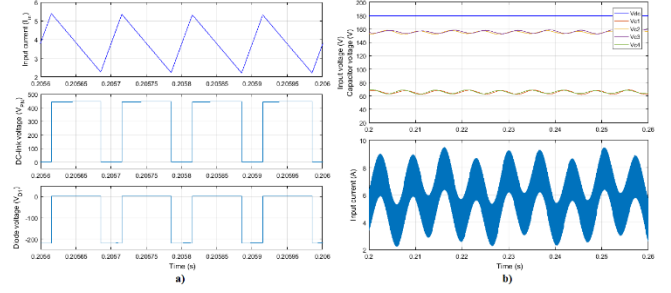


Fig. 5. Simulation results of qZST<sup>2</sup>I under normal mode. a) input current ( $I_{in}$ ) DC-link voltage ( $V_{PN}$ ), diode voltage ( $V_{D1}$ ). b) input voltage ( $V_{ik}$ ), capacitor voltage ( $V_{C1}$ ,  $V_{C2}$ ,  $V_{C3}$ ,  $V_{C4}$ ), input current ( $I_{in}$ ).

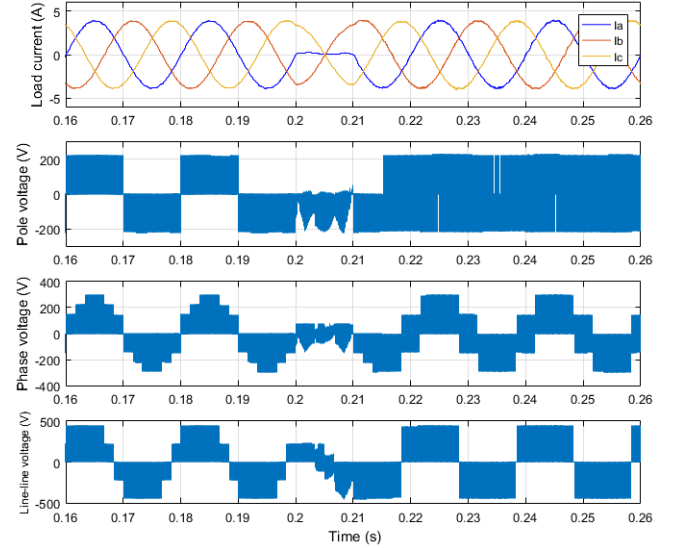


Fig. 6. Simulation results of qZST<sup>2</sup>I under normal and fault mode of the  $S_{1a}$ . From top to bottom: Output load current ( $I_A$ ,  $I_B$ ,  $I_C$ ), output pole voltage ( $V_{AO}$ ), output phase voltage ( $V_{AG}$ ), output line-line voltage ( $V_{AB}$ ).

### III. SIMULATION RESULTS

The efficiency of the proposed qZST<sup>2</sup>I is verified with the help of MATLAB Simulink software. The parameters using in simulation are listed in Table II. When applying 180 V at DC input voltage, the ST duty ratio  $D_0$  and modulation index are respectively set to 0.3 and 0.7 to achieve 110 V<sub>RMS</sub> at output voltage.

The simulation results of the proposed configuration are shown in Fig. 5. The input current ( $I_{in}$ ), which is also the inductor current ( $I_{L1}$ ), is increased in ST mode which is represented in time interval that the DC-link voltage ( $V_{PN}$ ) is zero and the diode ( $D_1$ ) is reserved bias, as presented in Fig. 5(a). By applying the ST duty ratio ( $D_0 = 0.3$ ) for the proposed topology, the average value of capacitor voltages

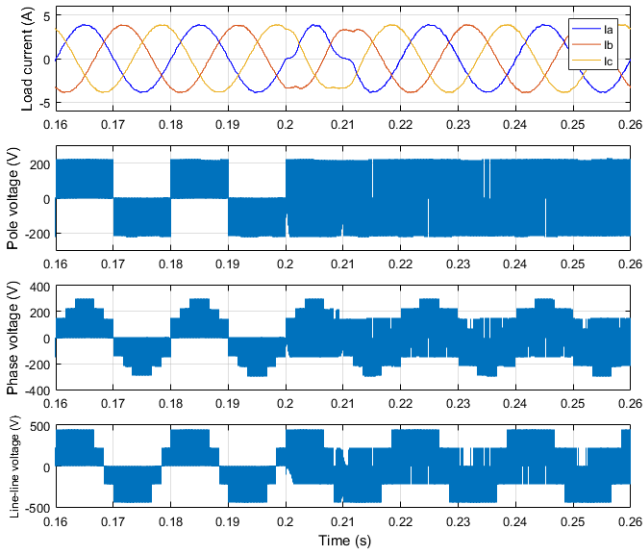


Fig. 7. Simulation results of qZST<sup>2</sup>I under normal and fault mode of the bi-directional switch ( $S_{2a}$  and/or  $S_{3a}$ ). From top to bottom output load current ( $I_a$ ,  $I_b$ ,  $I_c$ ), output pole voltage ( $V_{Ao}$ ), output phase voltage ( $V_{AG}$ ), output line-line voltage ( $V_{AB}$ ).

( $V_{C1}$ ,  $V_{C2}$ ,  $V_{C3}$ ,  $V_{C4}$ ) are 66.6V, 157V, 157V and 66.6V, respectively, as illustrated in Fig. 5(b). As a result, the peak value of DC-link voltage is approximately 450V. The average value of input current is 5.82A which is presented in Fig. 5(b).

By using 0.7 for modulation indices, the RMS value of output load voltage is 110V<sub>RMS</sub>. Thus, the RMS value of output load current is 2.74A<sub>RMS</sub>. When the OC fault occurring at  $S_{1a}$ , there are no the value  $+V_{PN}/2$  at output pole voltage, as shown in Fig. 6. Therefore, the output load current is maintained zero during half positive cycle of load current. By applying the introduced PWM technique, the converter behaves like a two-level boost inverter and therefore, the operation of the system is restored, as illustrated in Fig. 6.

Not serious than the  $S_{1a}$  OC fault, the bi-directional OC fault just generates the distortion at output load voltage and current, so the quality of output is decreased, as shown in Fig. 7. By changing the control signal of faulty phase to two-level inverter, the operation of the converter is maintained, as illustrated in Fig. 7.

#### IV. CONCLUSIONS

This paper has presented a redundant unit form of three-phase qZS three-level T-type inverter. This combination provides a fault-tolerant capability which guarantees the operation of the converter under open-circuit fault of switching devices. Therefore, the reliability of the converter is significantly improved. The operating principle and theoretical analysis were presented in detail and validated by using MATLAB Simulink software.

#### ACKNOWLEDGMENT

This research was funded by CT.2019.04.03 project. This work was supported by the Advanced Power Electronics Laboratory, D405 at Ho Chi Minh City University of Technology and Education, Viet Nam.

#### REFERENCES

[1] Siddhartha A. Singh, Giampaolo Carli, Najath A. Azeez, Sheldon S. Williamson, "Modeling, Design, Control, and Implementation of a Modified Z-Source Integrated PV/Grid/EV DC Charger/Inverter",

IEEE Transactions on Industrial Electronics, vol. 65, no. 6, pp. 5213-5220, 2018.

- [2] A.H. Rajaei, M. Mohamadian, S.M. Dehghan, A. Yazdian, "Single-phase induction motor drive system using z-source inverter", IET Electric Power Applications, vol. 4, no. 1, pp. 17-25, 2010.
- [3] Fang Zheng Peng, "Z-source inverter", IEEE Transactions on Industry Applications, vol. 39, no. 2, pp. 504-510, 2003.
- [4] Fang Zheng Peng, Miaosen Shen, Zhaoming Qian, "Maximum boost control of the Z-source inverter", IEEE Transactions on Power Electronics, vol. 20, no. 4, 2005.
- [5] Yu Tang, Shaojun Xie, Jiudong Ding, "Pulsewidth Modulation of Z-Source Inverters With Minimum Inductor Current Ripple", IEEE Transactions on Industrial Electronics, vol. 61, no. 1, pp. 98-106, 2014.
- [6] Poh Chiang Loh, Feng Gao, Frede Blaabjerg, Shi Yun Charmaine Feng, Kong Ngai Jamies Soon, "Pulsewidth-Modulated Z-Source Neutral-Point-Clamped Inverter", IEEE Transactions on Industry Applications, vol. 43, no. 5, pp. 1295-1308, 2007.
- [7] Poh Chiang Loh, Sok Wei Lim, Feng Gao, Frede Blaabjerg, "Three-Level Z-Source Inverters Using a Single LC Impedance Network", IEEE Transactions on Power Electronics, vol. 22, no. 2, pp. 706-711, 2007.
- [8] Xiaoquan Zhu, Bo Zhang, Dongyuan Qiu, "A High Boost Active Switched Quasi-Z-Source Inverter With Low Input Current Ripple", IEEE Transactions on Industrial Informatics, vol. 15, no. 9, pp. 5341-5354, 2019.
- [9] Joel Anderson, F.Z. Peng, "Four quasi-Z-Source inverters", 2008 IEEE Power Electronics Specialists Conference, Jun. 2008, pp. 2743-2749.
- [10] Changwei Qin, Chenghui Zhang, Alian Chen, Xiangyang Xing, Guangxian Zhang, "A Space Vector Modulation Scheme of the Quasi-Z-Source Three-Level T-Type Inverter for Common-Mode Voltage Reduction", IEEE Transactions on Industrial Electronics, vol. 65, no.10, pp. 8340-8350, 2018.
- [11] Deqing Yu, Qiming Cheng, Jie Gao, Fengren Tan, Yu Zhang, "Three-level neutral-point-clamped quasi-Z-source inverter with reduced Z-source capacitor voltage", Electronics Letters, vol. 53, no. 3, pp. 185-187, 2017.
- [12] Mokhtar Yaghoubi, Javad Shokrollahi Moghani, Negar Noroozi, Mohammad Reza Zolghadri, "IGBT Open-Circuit Fault Diagnosis in a Quasi-Z-Source Inverter", IEEE Transactions on Industrial Electronics, vol. 66, no. 4, pp. 2847-2856, 2019.
- [13] V. Fernão Pires, Armando Cordeiro, Daniel Foito, Joao F. Martins, "Quasi-Z-Source Inverter With a T-Type Converter in Normal and Failure Mode", IEEE Transactions on Power Electronics, vol. 31, no. 11, pp. 7462-7470, 2016.
- [14] Borong Wang, Zhan Li, Zhihong Bai, Philip T. Krein, Hao Ma, "A Redundant Unit to Form T-Type Three-Level Inverters Tolerant of IGBT Open-Circuit Faults in Multiple Legs", IEEE Transactions on Power Electronics, vol. 35, no. 1, pp. 924-939, 2020.



# A High-Efficient Power Converter for Thermoelectric Energy Harvesting

Van-Khoa Pham

Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education

Ho Chi Minh City, Vietnam  
khoapv@hcmute.edu.vn

**Abstract**—Self-power technique is a vital key for stand-alone applications whereas battery replacement may be impossible. For wearable applications, extracting energy from the ambient temperature is one of the best solutions among the other energy harvesting methods such as solar, wireless waves, and temperature. In this paper, a high-efficient power dc-dc converter with maximum power point tracking (MPPT) and zero-current switching (ZCS) based on digital counters is proposed for thermoelectric energy harvesting. The proposed technique is able to adapt to a wide range of temperature differences. The integrated ZCS module plays an essential role in reducing the loss induced by inaccurately controlling the high-side switch. Besides, the maximum power extracted from the thermal energy source is monitored with the MPPT module. The power converter was simulated using CMOS 600nm Nuvoton technology. From the simulation results, it shows that when employing a thermoelectric generator with a temperature gradient of 3 Celsius degrees, the converter is capable of providing a maximum power of 112 $\mu$ W with a high-efficient of 66%.

**Keywords**—thermoelectric generator (TEG), dc-dc booster, discontinuous conduction mode, zero-current switching (ZCS), maximum power point tracking (MPPT), counter-based controller

## I. INTRODUCTION

Success in semiconductor and sensing technology has strongly supported the internet of things (IoT) concept in which a large number of devices can be wirelessly connected and communicated to each other. Wearable devices designed for specific applications have successfully applied the IoT concept. Ultra-low power consumption is a vital key in wearable applications because of the size constraints for batteries [1]. Even though circuit designs have been significantly optimized to consume as little power as possible [1]. However, battery capacity is determined by how long a wearable product can be operated properly [2]. For on-body applications, energy harvesting techniques from the ambient environments such as solar, wireless waves, and temperature

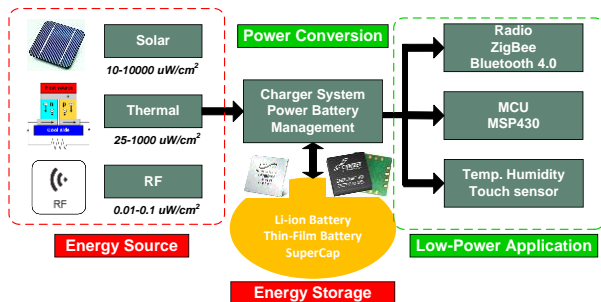


Fig. 1. Energy harvesting systems for wearable applications

have gained a lot of attraction from the circuit design community because the harvested power from renewable

resources can substitute for conventional power delivery methods [3], [4], [5]. As demonstrated in Fig. 1, an energy harvesting system consists of three main parts: energy sources, power conversion, and energy storage devices. The power conversion plays an essential role in providing a high-efficiency output power for ultra-low-power applications [1-6].

In comparison to the other energy harvesting sources mentioned previously, thermoelectric generators (TEGs) seem to be suitable for wearable applications because the harvested power is only dependent on temperature conditions [6-9]. As depicted in Fig. 2a, a thermoelectric generator is constructed by a lot of thermocouple elements which are connected electrically in series. Based on the Seebeck effect [10], the output voltage from TEGs is proportional to the temperature gradient between the hot and cold sides. For the wearable applications, the hot and cold sides can be the surface of the skin and ambient environment, respectively [11]. It should be noted that the voltage level harvested from TEG devices is not high enough to power any digital circuitry. Therefore, a dc-dc converter is very essential to boost the low-voltage level generated by TEGs up to a useful voltage level for powering CMOS operations [1], [2].

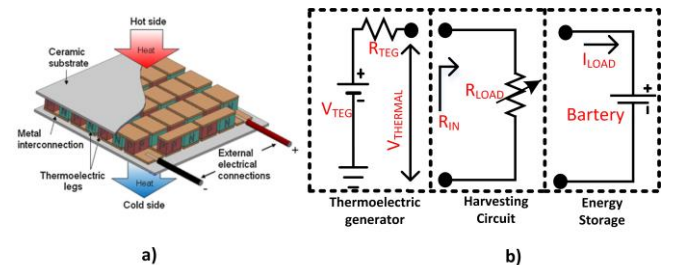


Fig. 2. a. Thermocouple structure [6] b. System overview of a thermoelectric energy harvesting system [3]

As shown in Fig. 2b, the system overview of a thermoelectric energy harvesting system consists of three parts: harvested source as TEGs, harvesting circuit, and energy storage. Where the magnitude of  $V_{TEG}$  reflects the temperature gradients applied on two sides of TEGs. The internal resistance ( $R_{TEG}$ ) depends on the specific structure of TEG. For wearable applications, the TEG products from [12] introduce a large internal resistance as large as 185 $\Omega$ . Here, the harvesting circuit is optimized to make the conversion ratio between the harvested power and the delivered power as high as possible.

$$Power\ eff.\ (\%) = \frac{P_{OUT}}{P_{IN}} \quad (1)$$

From Equ. 1, it is obvious that we are able to obtain a high power efficiency when maximizing the delivery power

generated from the dc-dc converter. Based on the conversation law, the summation of the output power, the power consumption of the controller, switching, and synchronization losses as well is equivalent to the input extracted power from TEGs [4]. Therefore, power efficiency can be improved significantly if the power losses in the energy harvesting system are taken into account [1], [9]. To do so, switching signals need to be controlled properly to minimize synchronization losses. Also, it is very essential to utilize pure and simple digital circuits in implementing the dc-dc boost controller. This study aims to investigate the impacts of switching powers and synchronization losses. We proposed a high efficient power dc-dc boost converter in which the switching signals for both the low-side and high-side switches are controlled accurately to extract maximum input power from TEGs and minimize synchronization losses. By doing so, the digital controller is designed mainly for using counters which consume very low power.

## II. PROPOSED ENERGY HARVESTING DESIGN

The proposed block diagram depicted in Fig. 3 consists of two main parts: the starter circuit and the main dc-dc boost converter. For on-body applications, the voltage output harvested from TEGs is extremely low and not sufficient for powering any CMOS circuitry. Therefore, the starter circuit is required in order to generate a higher voltage to start the main converter manually for the first time. This study aims to demonstrate the startup operation from 60mV. As shown in Fig. 3, the energy harvesting circuit is connected directly to the TEGs. After the starting process, the output voltage of the starter circuit,  $V_{DDS}$ , reaches 1.5. This voltage level is high enough to power the proposed digital MPPT-ZCS controller. The goal of the proposed design is to obtain a high voltage level as 4.2 V for charging energy storage devices such as big capacitors or batteries. In this way, wearable products can self-sustain its operation without external power sources.

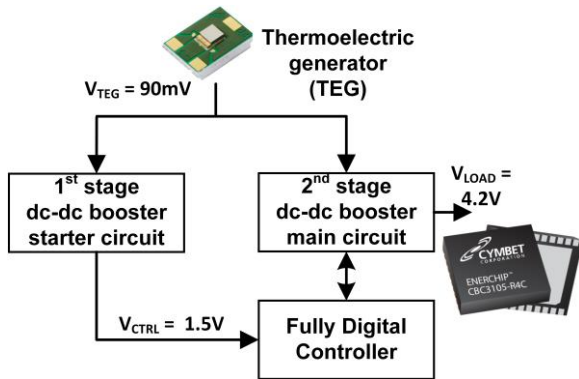


Fig. 3. The block diagram of the proposed thermoelectric energy harvesting system

According to the block diagram of the proposed thermoelectric energy harvesting system, the detailed schematic of the proposed system is demonstrated in Fig. 4. It includes two main parts: the starter circuit and main dc-dc boost converter in which the switching signals are controlled properly by the MPPT-ZSC controller. The input voltage node,  $V_{in}$ , for the converters is connected directly to single or array thermoelectric generators which are modeled by a voltage source  $V_{TEG}$  combined with a serial resistance  $R_{TEG}$ . Generally, the large capacitor,  $C_{in}$ , is needed in reducing the

input-voltage ripple. Depicted on Fig. 4, for the starter circuit, the on/off switch,  $S_1$ , linked with a serial resistor,  $R_{SW}$  is to illustrate a real mechanical switch that is activated by human. The switch will be pressed manually to start the main dc-dc boost converter. The diode,  $D$ , acted as the high side switch is made by a diode-connected transistor. The startup operation is explained as follows. If the switch  $S_1$  is pressed manually, the harvested current from TEGs will flow through the boost inductor  $L$ . After that, when the switch  $S_1$  is un-pressed, the energy stored on the inductor  $L$  will be released via the diode  $D$  in order to charge the output capacitor,  $C_{DDS}$ . By doing so, we can obtain the startup voltage as high as 1.5V when the energy flows into the capacitor  $C_{DDS}$ . This level voltage is helpful to power the operation of the main dc-dc booster. The high-efficient boost converter proposed in this study is based on conventional boost converters in which the high and low-side switches in discontinuous conduction mode are controlled properly. As shown in Fig. 4, the transistors,  $M_1$  and  $M_2$ , are used to demonstrate the low-side and high-side switches, respectively. One thing that should be noted here is that the high-side switch,  $M_2$ , is essential for the boost converter to reduce significantly the power losses suffered by both the voltage drop and the synchronization, which will be explained in detail later.

The purpose of the MPPT-ZCS controller is to monitor the extracted power from TEGs and then adjust accurately the on-time for the switching signals on both the low-side and high-side switches using the digital switching signal QN and QP, respectively. The signal MPP is used to check periodically the average input voltage for the maximum power point tracking operation. Conceptually, we can obtain the maximum extracted power if the input voltage,  $V_{IN}$ , is maintained at half of  $V_{TEG}$ .

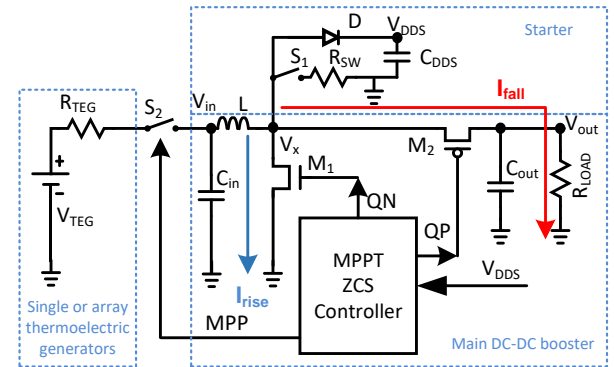


Fig. 4. The schematic of the proposed thermoelectric energy harvesting system

The waveforms in controlling both the low-side and high-side switches are presented in Fig. 5. As previously mentioned, the discontinuous conduction mode (DCM) applied to the boost converter for achieving a high-efficient conversion ratio is superior to the continuous conduction mode (CCM) [4]. By occupying DCM and an efficient control method for the main dc-dc boost converter, we can prevent the flowing negative current discharge the output capacitor  $C_{OUT}$  as depicted in Fig. 4. The previous studies [1], [2], [4] show that the negative current is the main reason to increase the synchronization power loss and switching loss as well. The waveforms in discontinuous-conduction mode shown by Fig. 5, in which  $T_1$  and  $T_2$  are the duty-adjustable



pulses for switching transistors  $M_1$  and  $M_2$ , respectively. Besides,  $I_{RISE}$  and  $I_{FALL}$  are used to denote the current flows through the low-side and high-side switches, respectively. Based on the operation of the DCM, the magnitude of  $I_{RISE}$  is proportional to the on-time of the low-side switch.

Here, the switching frequency for boost converters should be considered carefully to optimize the conversion ratio. Equ. 2 shows the impact of the values for  $R_{TEG}$  and the induction  $L$  on the boost converter to the switching frequency [2].

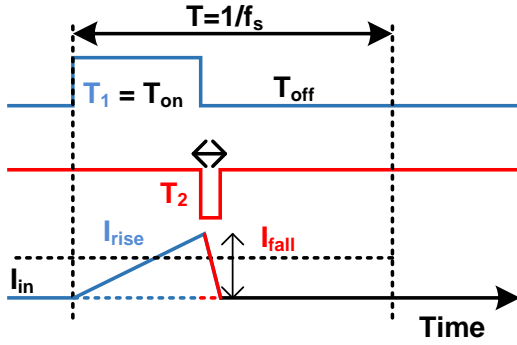


Fig. 5. The waveforms in discontinuous-conduction mode

$$f_s = \frac{R_{TEG}}{8L} \quad (2)$$

The prior study [9] claimed that a higher inductance can reduce the power losses significantly. Therefore, we used  $100\mu H$  for the boost inductor  $L$  in this work because of the limitation of form-factor and a small equivalent series resistance as well. In addition, the thin-film thermoelectric generator used for our simulations [12] introduces a large internal resistance as  $185\Omega$ . If we use a duty cycle of 50 in controlling the low-side switch  $M_1$ . As a result, a very high-frequency in this case is required as calculated by Equ. 2. Even though the proposed converter is connected to an array of TEGs. If so, this results in an inefficient power converter. Because the experimental results [1] show that a high switching frequency of boost converters is not suitable in optimizing the switching power loss as well as power consumption. One more thing should be noted here, when using simple circuits, it is not easy to generate a high-frequency signal precisely to obtain an expected on-time level for the low-side switch  $M_1$ . Therefore, instead of changing the switching frequency, we realized a frequency as low as 20KHz combined with a duty-adjustable method for controlling the low-side switch  $M_1$  in this work. Its operation will be explained later in this paper.

$$T_1 = \sqrt{\frac{2L}{R_{IN} * f_s}} \quad (3)$$

Once both the switching frequency, the boost inductor, and the internal resistance value of TEGs are given, the on-time,  $T_1$ , for the low-side switch  $M_1$  should be controlled accurately as presented by Equ. 3 to implement the maximum power point tracking function.

### III. CIRCUIT IMPLEMENTATION

Fig. 6 shows conceptually the block diagram of the proposed fully digital controller including the MPPT and ZCS controller for the low-side and high-side switch, respectively. As mentioned above, to extract the maximum input power

from TEGs, the dc-dc boost converter controls its internal impedance ( $R_{IN}$ ) by adjusting the pulse width of  $T_1$  to match  $R_{IN}$  with the  $R_{TEG}$ . To do so, a counter-based MPPT controller shown in Fig. 6a is proposed. It comprises two main parts: the voltage sensor and the duty controller. The voltage sensor is constructed by a low pass filter, voltage divider, and voltage comparator as well. The low pass filter is essential to have an average input voltage from TEGs. The input of the voltage divider is connected directly to the output of TEGs. The fact that the relationship between maximum power voltage point ( $V_{MPP}$ ) and the open-circuit voltage ( $V_{TEG}$ ) is almost linear [13]. Therefore, for thermoelectric energy harvesting applications,  $V_{MPP}$  should be calculated by the fractional open-circuit voltage as demonstrated in Equ. 4. By doing so, the counter-based MPPT controller will periodically isolate the thermoelectric generators with the boost converter and then measure the open-circuit voltage to determine the internal resistance of the converter is equivalent to the internal resistance or not. When the signal MPP is enabled, the outputs from both the low-pass filter and the voltage divider are applied to the voltage comparator. The comparison result is a binary level corresponding to the average input voltage level ( $V_{IN}$ ) and half of  $V_{TEG}$ . If  $V_{AVG} < V_{TEG}/2$ , the comparator output is 0. On the contrary, the output signal will be 1 when  $V_{AVG} > V_{TEG}/2$ . After that, the comparison result synchronized by the MPPT\_clk signal will determine whether or not the on-time pulse of the signal  $T_1$  for the low-side switch should be increased.

$$P_{IN,MAX} = \frac{V_{TEG}}{2} * \frac{V_{TEG}}{2R_{TEG}} = \frac{V_{TEG}^2}{4R_{TEG}} \quad (4)$$

As mentioned previously, a 20KHz of frequency is utilized for switching the low-side switch  $M_1$ . According to the method proposed in [14], we used a tapped delay line in digital pulse width modulation (DPWM) block formed by a ring oscillator combined with a 4-bit up-down counter to adjust the pulse width of the switching frequency precisely. Delay elements formed by resistors and capacitors in the delay line are calculated carefully to have the accurate step size for adjusting the pulse width. By doing so, the DPWM in this study can adjust the pulse width with a resolution as small as  $1.1\mu s$  to control the low-side switch  $M_1$ . As a result, the MPPT controller is able to cover a wide range of input voltage corresponding to a wide range of temperature changes. One thing should be noted here, the switch  $M_1$  is made by a large-size power transistor. Therefore, the digital internal signal QN\_D should be driven by a large-size buffer as shown in Fig. 6a.

To minimize the synchronization loss, the zero-current switching of the current stored on the  $L$  should be taken into account by controlling on-time for the high-side switch  $M_2$  accurately. According to the conversation ratio between the input harvested voltage from TEGs ( $V_{IN}$ ) and the boosted voltage ( $V_{OUT}$ ) as shown in Fig. 4, the time for opening the switch  $M_2$  can be calculated as the following equation

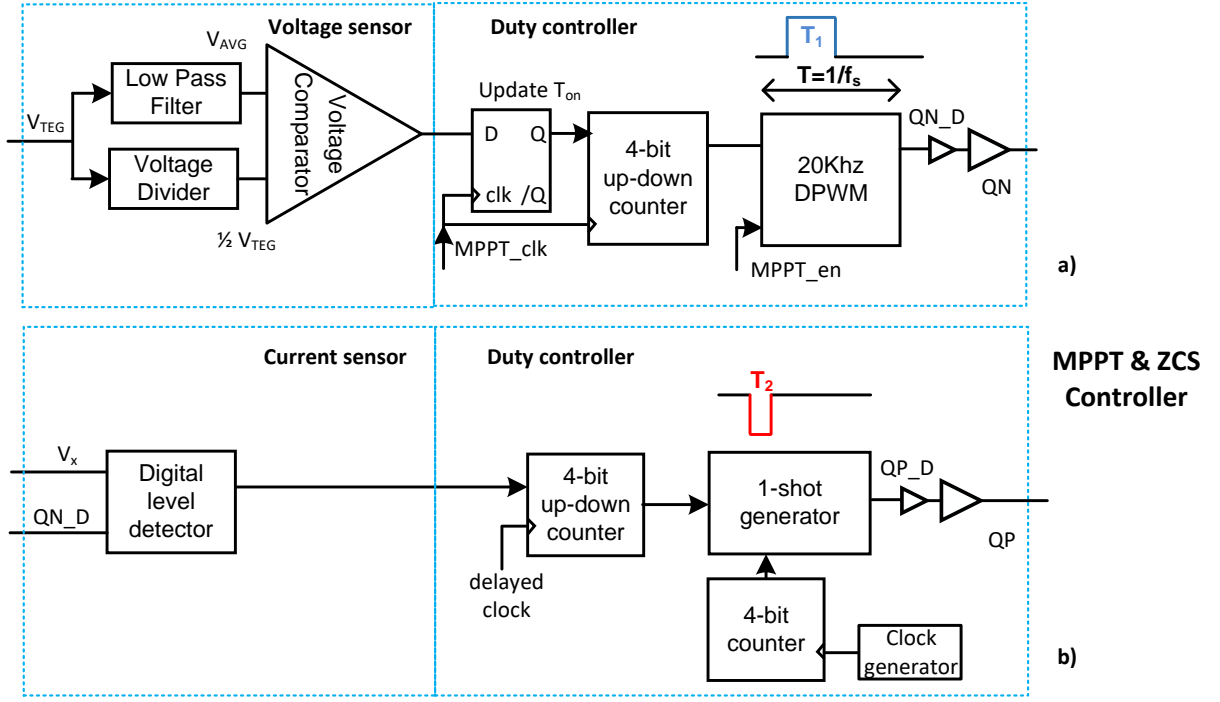


Fig. 6. The block diagram of the proposed fully digital controller a. counter-based MPPT controller and b. counter-based ZCS controller

$$\frac{T_2}{T_1} = \frac{V_{IN}}{V_{OUT} - V_{IN}} \quad (5)$$

As shown in Fig. 6b, a digital voltage level formed flip-flops is utilized to detect either the larger or a smaller voltage  $V_x$  in comparison to the output voltage level. According to the comparison result, if the high-side switch  $M_2$  is controlled accurately, the latency allowing the inductor current to flow negative can be reduced. The input of the digital level detector includes the voltage level from the  $V_x$  node and the internal signal for controlling the low-side switch  $M_1$ . Based on the output of the flip-flop clocked by the  $QN\_D$  signal, it will be defined as a high or low level. This signal is used to determine whether the 4-bit up/down counter in the proposed duty controller is increase or decrease values.

As described previously, for self-sustain wearable applications, we expect to obtain an output voltage as high as 4.2V for charging storage components like batteries. From Equ. 5, we can see that the conversation ratio in this study is as high as 35 if the harvested voltage of TEGs is 120mV at 2 Celsius degree temperature gradient. If the pulse width for opening the low-side switch  $M_1$  is adjusted from 10 $\mu$ S to 25 $\mu$ S. As a result, the pulse width for the switch  $M_2$  is varied from 300nS to 735nS. By using the 4-bit counter as shown in Fig. 6b, the pulse width resolution for the switch  $M_2$  is around 27nS. From this analysis, the frequency for the clock generator depicted in Fig. 6b is defined to drive the 4-bit counter. The 1-shot generator constructed by a digital comparator will generate a pulse width  $QP\_D$  for opening the switch  $M_2$  properly based on the input counter values. In the same way with the signal  $QN\_D$ , a driver is needed for the signal  $QP\_D$  to drive a significant load as shown in Fig. 6b.

#### IV. SIMULATION RESULTS

The operation of the proposed converter was verified using CMOS 600nm Nuvoton technology [15] and the Spectre Circuit Simulator [16]. Table I illustrates the design parameters of the converter. The voltage level for controlling

the main dc-dc booster obtained from the startup process is defined as 1.5V. The internal resistance of TEGs is equivalent to 61 $\Omega$  when three TEG devices fabricated by Mircopelt [12] are linked in parallel. As mentioned previously, the specification of the TEG [12] was modeled by a voltage source connected in series with an internal resistor using Verilog-A language [17] as shown in Fig. 4. For the boost circuit, the inductor value in this work is as high as 100 $\mu$ H to minimize the conduction loss and maintain the startup process. For the wearable applications such as a thermal harvester on the human body presented in [11], [13], it is not easy to make a large temperature gradient between the ambient environment and the body heat due to imperfect issues involving the mechanical design. Therefore, in this work, we evaluated the operation of the proposed boost converter when the input voltage from TEGs is changed from 60mV to 300mV. The limit of the voltage range corresponds to the 1 to 5 Celsius degree of the temperature difference applied to the TEGs. The output voltage of the dc-dc boost converter is expected as high as 4.2V for normal operation.

TABLE I. DESIGN PARAMETERS FOR THE ENERGY HARVESTING CIRCUIT

Specification	Value
CMOS Technology	Nuvoton 600nm [15]
Seebeck voltage	60mV/ $^{\circ}$ C [12]
Boost inductor	100 $\mu$ H
Impedance matching	61 $\Omega$
Control voltage	1.5V
Input voltage range	60 - 300mV
Output voltage	4.2V

The Fig. 7 shows the simulated results for the proposed thermoelectric energy harvesting system with the different applied temperatures on TEGs. The graph demonstrates the relationship between the power efficiency, the delivered power, and the applied temperatures. In this works, we assumed that the maximum temperature difference is as high as 5 Celsius degrees corresponding to 300mV of the harvested input voltage. As we can see that the proposed converter is able to obtain a high percentage of power efficiency as 64% when the input voltage is as low as 90mV. By performing the same simulation conditions with 3 Celsius degrees of the applied temperature, the converter can generate 112μW of the output power corresponding to 66% of the power efficiency. When the harvested input voltage is as low as 60mV with only 1 Celsius degree difference, the proposed converter is still operating. However, if the input voltage is lower than 60mV, the harvesting circuit will be halted because of low  $V_{DDS}$ . In this case, the converter requires a manual process to start the system again.

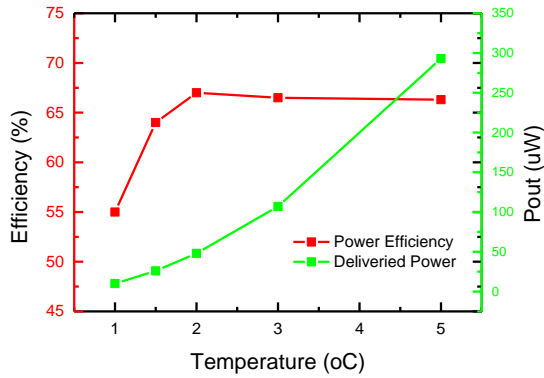


Fig. 7. power efficiency of the proposed thermoelectric energy harvesting system with the different applied temperatures

In addition, to demonstrate the impact of the operation of the maximum power point tracking module, we did a performance comparison between the harvesting circuit using MPPT and without the MPPT function. To do so, we use an enable signal MPPT\_en from an external switch to define whether the converter is without or with the MPPT operation. As shown by Table II, it is obvious that the harvesting circuit combined with the proposed MPPT controller is able to tune the pulse width below 1μs in controlling the low-side switch  $M_1$ . The proposed circuit having the MPPT mechanism can achieve a higher power efficiency compared with non-MPPT at 180mV of the input voltage. The power efficiency difference is as large as 23% when the low-side switch  $M_1$  is controlled by an inaccurate pulse width signal.

TABLE II. POWER EFFICIENCY COMPARISON USING THE MPPT APPROACH

Specification ( $V_{TEG}=180mV$ )	$T_1$ (μS)	Power Efficiency (%)
MPPT	15.4	66
non-MPPT	13	43

## V. CONCLUSION

In this study, we proposed a high-efficient power converter for wearable thermoelectric energy harvesting applications. A fully digital controller was designed to consider carefully power losses suffered by the un-matching internal resistance of the power converter when half of the harvested voltage from TEGs is not maintained as the input voltage as well as discharging the output capacitor when the high-side switch is not controlled accurately. The idea was verified using CMOS 600nm Nuvoton technology. From the simulation results, the proposed dc-dc boost converter can deal with a wide range of temperature changes. The power efficiency is maintained as high as 64% even though the harvested input voltage is as low as 90mV.

## ACKNOWLEDGMENT

This work belongs to the project in 2020 funded by Ho Chi Minh City University of Technology and Education, Vietnam.

## REFERENCES

- [1] S. Bandyopadhyay, P. P. Mercier, A. C. Lysaght, K. M. Stankovic, and A. P. Chandrakasan, "A 1.1 nW Energy-Harvesting System with 544 pW Quiescent Power for Next-Generation Implants," *IEEE Journal of Solid-State Circuits*, vol. 49, pp. 2812 - 2824, 2014.
- [2] Y. K. Ramadass and A. P. Chandrakasan, "A Battery-Less Thermoelectric Energy Harvesting Interface Circuit With 35 mV Startup Voltage," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 333 - 341, 2010.
- [3] S. Bandyopadhyay and A. P. Chandrakasan, "Platform Architecture for Solar, Thermal, and Vibration Energy Combining With MPPT and Single Inductor," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 2199 - 2215, 2012.
- [4] E. J. Carlson, K. Strunz, and B. P. Otis, "A 20 mV Input Boost Converter With Efficient Digital Control for Thermoelectric Energy Harvesting," *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 741 - 750, 2010.
- [5] Khoa Van Pham, etc., "A Thermoelectric Energy Harvesting Circuit For a Wearable Application," *Institute of Korean Electrical and Electronics Engineers*, vol. 21, iss. 1, pp. 66-69, 2017.
- [6] S.B. Riffat and X. Ma, "Thermoelectrics: a review of present and potential applications," *Applied Thermal Engineering*, vol. 23, no. 8, pp. 913-935, Jun. 2003.
- [7] M. Thielena, L. Sigrisb, M. Magno, C. Hierolda, and L. Benini, "Human body heat for powering wearable devices: From thermal energy to application," *Science Direct*, vol. 131, pp. 44-54, 2017.
- [8] J. Kim and C. Kim, "A DC-DC Boost Converter With Variation-Tolerant MPPT Technique and Efficient ZCS Circuit for Thermoelectric Energy Harvesting Applications," *IEEE Transactions on Power Electronics*, vol. 28, pp. 3827-3833, 2013.
- [9] J. Katic, S. Rodriguez, and A. Rusu, "A Dual-Output Thermoelectric Energy Harvesting Interface With 86.6% Peak Efficiency at 30 μW and Total Control Power of 160 nW," *IEEE Journal of Solid-State Circuits*, vol. 51, pp. 1928 - 1937, 2016.
- [10] S. Dalola, M. Ferrari, V. Ferrari, and M. Guizzetti, "Characterization of Thermoelectric Modules for Powering Autonomous Sensors," *IEEE Transactions on Instrumentation and Measurement* 2008.
- [11] A. Myers, R. Hodges, and J. S. Jur, "Human and environmental analysis of wearable thermal energy harvesting," *Energy Conversion and Management*, vol. 143, pp. 218-226, 2017.
- [12] TGP-651 Thin Film Thermogenerator, [http://www.micropelt.com/fileadmin/user\\_upload/\\_PDF\\_TGP\\_UK.pdf](http://www.micropelt.com/fileadmin/user_upload/_PDF_TGP_UK.pdf).

- [13] A. Paraskevas and E. Koutroulis, "A simple maximum power point tracker for thermoelectric generators," *Elsevier Energy Conversion and Management*, vol. 108, pp. 355-365, 2015.
- [14] A. Syed, E. Ahmed, D. Maksimovic, and E. Alarcon, "Digital pulse width modulator architectures," *Power Electronics Specialists Conference*, 2004.
- [15] <https://www.nuvoton.com/>
- [16] [https://www.cadence.com/en\\_US/home/tools/custom-ic-analog-rf-design/circuit-simulation/spectre-simulation-platform.html](https://www.cadence.com/en_US/home/tools/custom-ic-analog-rf-design/circuit-simulation/spectre-simulation-platform.html)
- [17] <https://literature.cdn.keysight.com/litweb/pdf/ads2004a/pdf/verilogaref.pdf>

# Collaborative Robotics in Construction: A Test Case on Screwing Gypsum Boards on Ceiling

Milan Gautam  
HAMK Tech

HAMK University of Applied Sciences  
Riihimäki, Finland  
milan.gautam@hamk.fi

Hannu Fagerlund  
HAMK Tech

HAMK University of Applied Sciences  
Riihimäki, Finland  
hannu.fagerlund@hamk.fi

Blerand Greicevci  
HAMK Tech

HAMK University of Applied Sciences  
Riihimäki, Finland  
blerand.greicevci@hamk.fi

Francois Christophe  
HAMK Tech

HAMK University of Applied Sciences  
Riihimäki, Finland  
francois.christophe@hamk.fi

Jarmo Havula  
HAMK Tech

HAMK University of Applied Sciences  
Riihimäki, Finland  
jarmo.havula@hamk.fi

**Abstract**— The use of collaborative robots (cobots) and human robot collaboration has started to increase in diverse industrial areas in the past years. However, cobots are seldom applied in construction work even if there is a growing need for robotics solutions in this area. This study aims to present the possibilities of applying lightweight cobots for single tasks in the construction sector. This paper presents the development of a proof-of-concept dedicated to screwing gypsum board panels to the ceiling of a room. After identifying the challenges related to this task, we present the cobot solution that we implemented for this purpose. Our results show that repetitive screwing work in upright position achieved by cobot can benefit workers by removing ergonomical strain. At the same time, cobot ensures that this task is achieved with constant performance. Future research directions lead to the search for better accuracy in screw positioning and more dynamic ways of collaboration with human workers.

**Keywords**— Collaborative robotics, Single-task construction robotics, Proof-of-concept design

## I. INTRODUCTION

Robotizing inner work in the construction field presents interest as the tasks are repetitive and monotonous but still currently in most cases carried out manually. Such inner work tasks, for example, ceiling surface treatment, painting, or mounting ceiling panels, require workers to change places after only a few minutes as stated in [1]. Despite the solutions developed in the early 21<sup>st</sup> century [2], their use is limited to large construction sites due to heaviness and high cost [3]. The rise of collaborative robots (Cobots) as the 2<sup>nd</sup> generation of robots opens new opportunities for developing single-task construction robots that work in collaboration with workers. This article reports on the possible use of cobots in the construction sector and the challenges that would need to be overcome to extend the use of cobots in this sector. The article presents an overview of the use of cobots in the construction sector in section II, continuing with the identified challenges and opportunities in section III. Section IV of the article describes the specific method used for the research case, following with the results in section V and the final conclusions in section VI.

## II. RELATED WORK: USE OF COBOTS IN CONSTRUCTION SECTOR

Our review is separated into two types of construction categories: outdoor applications (presented in II-A) and indoor applications (presented in II-B).

### A. Outdoor Applications

Construction automation of outdoor activities are being researched since 1970s in various countries like Japan, US, and Germany [2]. At that time, the initial idea is to build a fully autonomous robotic system that can provide the same benefit as robots inside factory floors running as mass manufacturing of cars. At present, due to the rapid development in Cobots technology, the idea of semi-automated man-machine collaborative systems is being proposed and implemented [4]. The recent examples are SAM and MULE135 robotic systems from Construction Robotics<sup>1</sup> and Tiger-stone robots for paving streets from Vanku B.V.<sup>2</sup>. One of the areas where collaborative robotics will tend to develop in outdoor construction is the additive manufacturing of building or building components utilizing multirobot systems together with autonomous mobile vehicles [5].

### B. Indoor Applications

The use of cobots in indoor construction environments is the subject of research interest all around the world. The ability of collaborative technology, whether it is a Cobot arm or a mobile robot that can autonomously navigate, has presented a wide range of opportunities for the construction industry [6]. Utilizing the technological advancement, several industrial and research-based solutions are proposed. The recent development in drilling tasks for inside construction environment using Cobots is proposed by a start-up<sup>3</sup>. They have patented the technology for mobile robotic drilling apparatus and a method for drilling ceilings and walls [7]. In 2018, Transforma robotics lab from Nanyang Technological University presented pictobot, a co-operative painting robot developed for interior finishing using a Cobot [8]. According to [8], Cobot can facilitate

<sup>1</sup> <https://www.construction-robotics.com/>

<sup>2</sup> <http://tiger-stone.nl/>

<sup>3</sup> <https://www.nlink.no/>



collaboration in construction and the development of human ingenious skills required to complete construction tasks.

This brief overview of works related to robotics used in construction sectors allows the identification of the challenges and opportunities presented in Section III during the completion of screwing gypsum boards on the ceiling with a Cobot, but these can be generalized to other indoor construction tasks.

### III. IDENTIFIED CHALLENGES AND OPPORTUNITIES

This section presents the challenges related to the task of screwing gypsum boards on the ceiling as presented in III-A, and challenges related to the use of a Cobot presented in III-B. Opportunities of using cobots in the construction field are presented in III-C.

#### A. Challenges related to the task

This subsection presents the identified specific challenges of the screwing gypsum task. These challenges are independent of the task being realised by a Cobot or not. The following challenges were identified:

##### Placing and holding the board before screwing it

The first challenge to the task is to place and hold the gypsum board to the ceiling. The gypsum board weighs between 9-12kg/m<sup>2</sup>. This means the lightweight robot which has only 5 kg of payload cannot lift or hold a gypsum board on its own. Construction workers use a gypsum board lifter to lift the panels and place it on the ceiling or walls<sup>4</sup>. In this experiment, a lifting panel was used.

##### Vertical position (opposite to gravity)

From [9], ceiling-hanging tasks were perceived as most stressful in terms of physical stress, fall potential, and being struck by or against an object.

##### Variations in wood hardness

The force that needs to be applied changes according to the place of screwing. For example, the hardness of the gypsum board is rather homogeneous, but wooden beams present heterogeneity in wood with some knots in place. Previous studies [10] concerning external forces in the human hand while working with tools show an average pressure of 250kPa (0,25N/mm<sup>2</sup>).

##### Intuitive placing the screws in the right place

When a human is performing any task with a power tool like screw driving, vision and sensation in human provides extended benefits. In the gypsum board screwing task, if a worker feels that the screw is going to knot in the wood, he will unscrew and take another screw to place it next to the initial place. The worker can also feel a resistance in his arm and evaluate the need to provide force to drive the screw to wood. Robotic arms do not have the intuition level of humans.

##### Navigating the entire room environment

The room in which gypsum boards are being placed on the ceiling needs to be free of movement and workers (human or robot) need to be careful while navigating around

this room for the safety of other workers but also for wall structures and other pillars.

#### B. Challenges related to Cobots

This subsection presents specific challenges related to using a Cobot for the task, which includes force limitation in joints, reachability of the screwing place, and obstacle recognition. These are described as follows:

##### Force required in Cobot's joint

The cobot in our use is a force limited robot, which means if a certain force threshold value is reached, the robot will go to a protective stop. After the state of protective stop, a robot controller must be restarted, and the robot program will begin from start. The UR robot has a force limited to 250N in general<sup>5</sup>. This makes use of force-limited cobots challenging as screw driving with power tools requires a certain amount of force[11].

##### Reaching all places that need to be screwed

The cobots are limited by the reachability and longer the reach of the robot, robot equipment's get heavier requiring immobile setup. The UR5 used in our case has a reach of 850mm when fully extended. The performance of robotic arms differs according to their position. The ceiling is generally a big workspace for any robotic arm to have such reachability.

##### Recognizing obstacles

Obstacles such as the holder of the gypsum board, persons, and potential pillars need to be recognized and avoided. Robotic arms alone do not have floor mobility, not even to recognize holders of the gypsum board, workers in the site, and other obstacles like are not recognized by the arm itself. This present a challenging situation of robotic use in such environment.

##### Availability of tools for Cobots

Human performs everyday manipulation tasks which could not be performed by any robot or brain in existence today[12]. Human hands can grab any tools and with human perception can use them in the given context. We can find human tools from simplified versions to complex ones that are being used in construction. However, if we talk about the robot tools, they need to be specifically designed for each purpose task and are task based. Hours of engineering time must be spent to develop robotic tools. For Cobots to be accepted, they need to work with the same tools than construction workers. Cobot investment is already high for companies that they would not want to invest on different tools than the ones that workers use.

#### C. Opportunities with cobots

This subsection presents the opportunities brought in by Cobots features and the rapid prototyping advantage for Cobots.

##### Cobots features

Cobots have the property to provide a safe environment for human workers to accomplish tasks in collaboration with the robot. As stated in [13], "*Collaborative industrial robots*

<sup>4</sup> <https://www.talhu.fi/tuotteet/tyomaateknikka/levyvaunut-ja-levyhiisit/>

<sup>5</sup> <https://www.universal-robots.com/products/>



are harmless to the human worker, affordable, and easy to use and program. More importantly, studies on human-robot collaboration have indicated that a better productivity at the workplace can be achieved through the collaboration of a human worker with a safe and flexible robot.” Programming of the robot is the final stage to give a robot a set of instructions to complete a task. The level of understanding of robot programming depends on the robotic knowledge of the user. The proficient user-like robotics engineer can program a robot in a way that the commands are not understandable for a worker. The main opportunity of Cobots lays in their easiness of programming so that they can be used by people without prior expertise in robotics.

### Rapid prototyping advantage

The introduction of lightweight robotics together with 3D printing technology provides the rapid prototyping advantage to test ideas and applications. If we combine the third element of scanning and reverse engineering, the whole process will help to convert tools usually made for human to a robotic end-effector for testing. This idea is presented further in Section IV.

## IV. METHOD

The scientific method applied for the development of this case study is action design. We have broken down our design method into the following phases as shown in Fig. 1. The processes presented in this figure are explained in their respective subsections.

### A. Identification of the possible areas of use of lightweight Robotics by involving the construction professionals

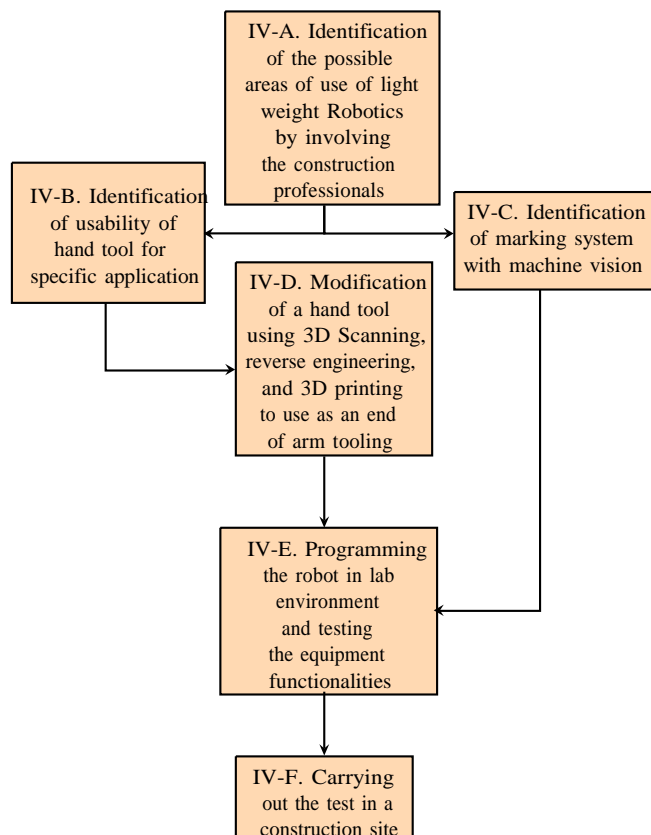


Fig.1 Block diagram representing different phases of design method

To gain an inner understanding of the construction problems, interviews were made for the possible cases together with Construction Engineering Lecturers, Construction Engineers, and construction workers. The economic considerations were the main point in creating solutions for different ideas. It was observed during the interview that these professionals can misinterpret the readiness of robotics technology and time to market estimation. Rapid ideation [14] was carried out and several prospective areas within construction and renovation were discovered. The ideas were generated in indoor construction tasks that are repetitive in nature and include injury-prone human ergonomics positions. One such task was overhead screwing of gypsum board panels. The installer will have to hold the tools and exert force to screw. Over the long working period on the same job, shoulder injuries were frequently reported by workers. The claim is validated by a scientific study on applied ergonomics [11]. Research from the start of the 21st century has shown that dry wall or gypsum board installation is one of the main injury-prone tasks in the construction industry [9]. Following the discussion and findings, the test was agreed on dry wall installation of gypsum board panels. The other agreement was that the robot should be able to use the hand tools that the human would use now.

### B. Identification of usability of hand tool for specific application

What if we could modify the hand tools and make them work with the robot? With this motive, some gypsum board screwing tools were looked and The Bosch GSR 18 V-EC +MA55 Cordless Screwdriver was selected for the experiment. The weight of the tool, easiness of modification was looked up on selection. The selection was also limited because of an available lightweight robot which supports a total payload of 5kg.

### C. Modification of a hand tool using 3D Scanning, reverse engineering, and 3D printing to use as an end of arm tooling

To fit a hand tool to a robotic arm, additional adapters or jigs are needed based on the geometry, reachability, and payload capacity of the robotic manipulators. To develop such adapters, information on the surface texture of the hand tool is required. There are 2 general methods to achieve the surface texture through 3D-model of the equipment readily available or 3D scanning tool and gaining digital data[15]. In this case, ATOS COMPACT scan<sup>6</sup> was used to digitize the tool surface texture. The 3D-CAD redesign was performed from the digital data. The adapter was designed based on the surface geometry of the tool as shown in Fig. 2, gained from the scanning tool. 3D printing of the adapter was carried out, as shown in Fig. 3 and tested to fit the tool geometry and attachment to a robotic flange.

<sup>6</sup> <https://www.gom.com/metrology-systems/atos/atos-compact-scan.html>



Fig. 2 Scanned model of the screwdriver

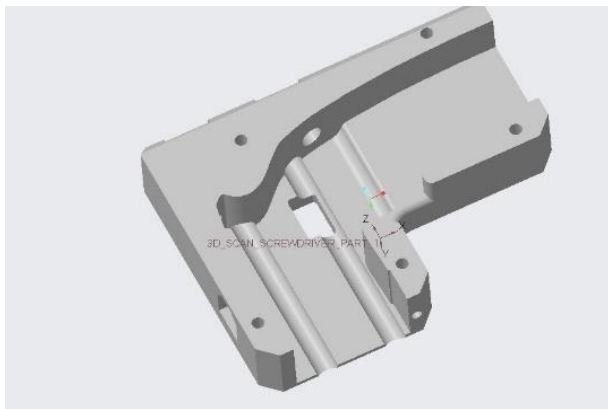


Fig. 3. Example of adapter CAD model design based on scanned model

The tool consists of turning off and on a switch, which was also needed to function automatically. For this, a pneumatic mechanism controlled with a solenoid valve was created. The screwdriver power supply was not modified in this case, so the battery power of the tool was used as shown in Fig.4. The additional elements required were 3D printed.

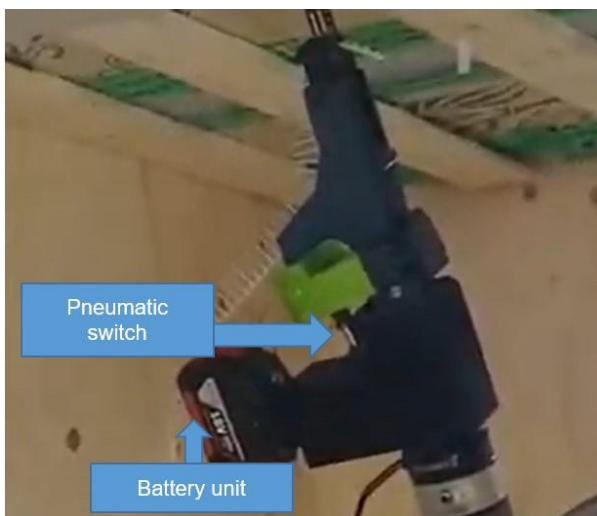


Fig. 4. Pneumatic switch developed for turning on and off the screwdriver

#### D. Identification of marking system with machine vision

Construction sites are unstructured environments where the human performs their task based on eye vision. Robots

need the same type of perception to screw at the right location and at the right distance. Machine vision systems have provided eyes to robots. To follow the simplistic approach, the simplest vision system that can be directly controlled with a robot controller was used<sup>7</sup>. It is not only important that the robot is equipped with a vision system but what objects or indicator it will recognize. In this case, target markings (see Fig.5) were trained to be recognized. These target markings are commonly used markers in the construction field.

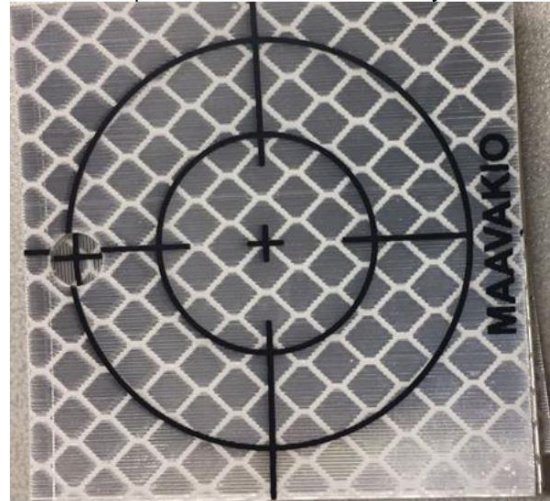


Fig. 5. Target Marking

#### E. Programming the robot in lab environment and testing the equipment functionalities

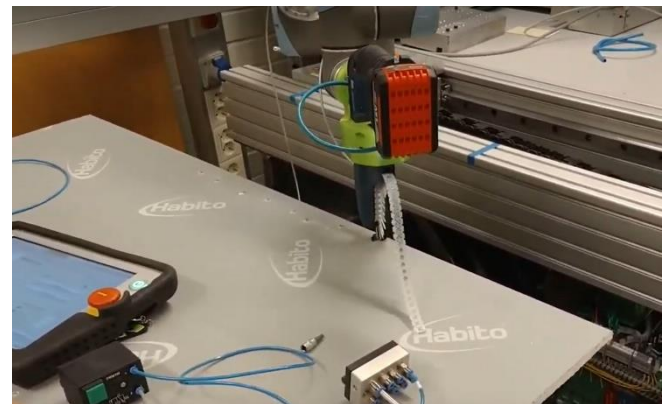


Fig. 6. Lab testing environment

Construction industry relies on their workers in the field, so if a robot is being used in the field alongside human workers, they should know how to operate it intuitively. Keeping this factor in mind, the robot was programmed in the lab simply using the commands available on the teach pendant of the robot. The programming was done so that users must place the robot in a location where the vision system can observe a marking easily and the robot is collision free. All equipment functionalities were also tested in the lab. The testing environment was flat for the screwing tasks in the lab as shown in Fig.6

<sup>7</sup> <https://robotiq.com/products/wrist-camera>

### F. Carrying out the test in a construction site

Tests were carried out in a wooden house construction site. The houses were being built in the construction engineering laboratory as shown in Fig.7



Fig. 7. Wooden House being built in Construction Lab

Two days of working time were reserved for testing. First day, all necessary equipment was made ready as shown in Fig.8 and the robot was reprogrammed to work inside the wooden house with allocated space. There was a need to reprogram the robot as the environment changed and for it not to hit the walls in such a closed environment. The next day, the experiment was demonstrated to the construction engineering lecturers, construction engineer, and students. Problems and challenges were observed. New ideas and suggestions were collected from the perspective of construction professionals.



Fig. 8. On site experimental setup

### V. RESULTS

Tests were conducted in two phases. For the first testing phase, the robot was screwing the screws systematically at regular intervals, whereas for the final testing phase each screwing position was previously inspected by the workers so that the screwing position could be avoided in case of a wooden eye at that position. When a wooden eye was noticed, the entire row of screwing coordinates was shifted to twice the size of the wooden eye to avoid the hardness of the wood in this place.

The first phase of testing was carried out with 2 rows of 3 screws and the trial was carried out for 13 times as can be seen from the linked video<sup>8</sup>. In total, 78 screws were inserted inside the gypsum board. During this test, 5 times failure occurred during the last screwing task and 8 times during the second last screwing position. When a failure occurred in the

second last position, the robot program needed to be restarted from the first screwing position, which means that the robot missed 2 screws. The total number of missed screws were 21 screws. The success rate in this first phase was approximately 78 percent if only considering how many screws went in. However, there were other variables like restarting time, adjusting the program to avoid previously screwed locations. According to the observation, the screwing task as a whole setup was effective around 30-40 percentage in this first phase.

The final test was fine-tuned avoiding all wooden eye locations for screwing in the robot program. In the final test, 3 rows of 3 screws per row were carried out. The total time to screw 9 screws was 113,64 seconds. The time for one screw installation was about 6,07 seconds. The camera detection time was 10,02 seconds and the rest of the time were to move from one screwing position to another.

From the experiment, it was observed that the gypsum board installation task can be carried out together with a collaborative robot and a construction worker working in the same workplace at the same time. The concept should be further developed and investigated to create a fully usable industrial solution.

### VI. CONCLUSION

The experiment presented here has shown the current possibilities and challenges that we have encountered during the screwing of a gypsum board to the roof with a collaborative robot.

The main findings from this project were the following:

- The robot goes to a protective stop if wooden eyes were present in the wood-to- screw gypsum board due to the hard wood material and safety limits of the robot.
- There might be a need of an additional force torque sensor to carry out the screwing precisely.
- The pneumatic switching should be replaced to direct power on and off for the screwing devices. Thus, the robot could control the switching directly.
- The feeding of the screw should be always consistent so that a precise contact between the screw head and the rotating screwing tool is made.
- The use of a collaborative robot will increase the ergonomics of the working of the construction workers.

Our future research in this area involves making the screwing tasks more dynamic with a mobile robot having a scissor lift. In this way, the combined cobot and mobile platform could cover the entire construction area. Additionally, it would be important to sense each screwing using the robot's force-torque sensor to notice the resistance to screwing due to a wooden eye. This would avoid that the robot would go into a protective stop.

### REFERENCES

<sup>8</sup> <https://youtu.be/CgTPEfGbLBI>



- [1] L. Koskela, H. Lempinen, and E. Salo, "The Feasibility of Construction Robotics in Finland and Norway," in *Proceedings of the 6th International Symposium on Automation and Robotics in Construction (ISARC)*, 1989, doi: 10.22260/isarc1989/0010.
- [2] T. Bock and T. Linner, *Construction Robots: Elementary Technologies and Single-task Construction Robots*. New York, NY: Cambridge University Press, 2016.
- [3] M. Pan, T. Linner, W. Pan, H. Cheng, and T. Bock, "A framework of indicators for assessing construction automation and robotics in the sustainability context," *J. Clean. Prod.*, vol. 182, pp. 82–95, 2018.
- [4] C. Han, "Human-robot cooperation technology an ideal midway solution heading toward the future of robotics and automation in construction," in *Proc. 28th Int. Symp. on Automation and Robotics in Construction, ISARC*, 2011, pp. 13–18.
- [5] H. Ardiny, S. Witwicki, and F. Mondada, "Construction automation with autonomous mobile robots: A review," in *2015 3rd RSI International Conference on Robotics and Mechatronics (ICROM)*, 2015, pp. 418–424.
- [6] J. M. D. Delgado *et al.*, "Robotics and automated systems in construction: Understanding industry-specific challenges for adoption," *J. Build. Eng.*, vol. 26, p. 100868, 2019.
- [7] H. Halvorsen, T. A. Henninge, and K. Fagertun, "Mobile robotic drilling apparatus and method for drilling ceilings and walls." Google Patents, 2018.
- [8] E. Asadi, B. Li, and I.-M. Chen, "Pictobot: a cooperative painting robot for interior finishing of industrial developments," *IEEE Robot. Autom. Mag.*, vol. 25, no. 2, pp. 82–94, 2018.
- [9] C. S. Pan, S. S. Chiou, H. Hsiao, J. T. Wassell, and P. R. Keane, "Assessment of perceived traumatic injury hazards during drywall hanging," *Int. J. Ind. Ergon.*, vol. 25, no. 1, pp. 29–37, 2000.
- [10] C. Hall, "External pressure at the hand during object handling and work with tools," *Int. J. Ind. Ergon.*, vol. 20, no. 3, pp. 191–206, 1997.
- [11] L. Yuan, B. Buchholz, L. Punnett, and D. Kriebel, "An integrated biomechanical modeling approach to the ergonomic evaluation of drywall installation," *Appl. Ergon.*, vol. 53, pp. 52–63, 2016.
- [12] C. C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments [grand challenges of robotics]," *IEEE Robot. Autom. Mag.*, vol. 14, no. 1, pp. 20–29, 2007.
- [13] K. Afsari, S. Gupta, M. Afkhamiaghda, and Z. Lu, "Applications of Collaborative Industrial Robots in Building Construction." August, 2018.
- [14] B. Clark and D. G. Reinertsen, "Rapid Ideation in Action: Getting Good Ideas Quickly and Cheaply," *Des. Manag. J. (Former Ser.)*, vol. 9, no. 4, pp. 47–52, 1998.
- [15] M. Sokovic and J. Kopac, "RE (reverse engineering) as necessary phase by rapid product development," *J. Mater. Process. Technol.*, vol. 175, no. 1–3, pp. 398–403, 2006.

# Impact of Financial Inclusion on Economic Growth: GMM Approach

Khac Hieu Nguyen

Department of Business Administration  
Ho Chi Minh City University of Technology and Education  
Vietnam  
hieunk@hcmute.edu.vn

Thi Anh Van Nguyen

Department of Business Administration  
Ho Chi Minh City University of Technology and Education  
Vietnam  
anhvan@hcmute.edu.vn

**Abstract** — This paper examines the influence of financial inclusion on economic growth in 37 developed countries and 21 emerging countries during the period 2006 – 2017. The GMM method is used to analyze panel data. The analysis results show that the financial inclusion has a positive effect on economic growth in developed and emerging countries. The impacts of financial inclusion on economic growth in developed countries has a steeper slope than in emerging countries. Besides financial inclusion, trade openness and intellectual property right also affect the economic growth of these countries.

**Keywords** — Financial Inclusion, Economic Growth, GMM.

## I. INTRODUCTION

Many countries today are interested in the term financial inclusion. Previous studies show that there is a relationship between financial inclusion and economic growth. Financial inclusion means the ease for all participants in an economy to access to credits, insurance and other formal financial services. High level of inclusivity in a financial society means that most of an economy's participants are using formal financial systems can get benefit from financial services. That is why financial inclusion is a topic that attracts researchers. However, the studies on this topic mainly focus on studying the factors affecting financial inclusion or study the effects of financial inclusion on economic growth. There are few research papers that focus on differences of financial inclusion among countries. This paper focus on differences of financial inclusion among countries especially the different impact of financial inclusion on economic growth. To investigate the relationship between financial inclusion and economic growth, we use GMM method to analyze the data because there is lag of dependent variables in the model. The research data are collected from Global Financial Development Database of World Bank. In the next section, we present the literature review of previous researches. Section 3 shows the research data and research method. The results are presented in section 4 and the last section is the conclusion.

## II. LITERATURE REVIEW

In late 1990s, the terminology “financial inclusion” received a great attention since as the policy-making issue of socially excluded people and research studies about the financial exclusion of socially excluded people have emerged [1]. After that, Leyshon & Thrift (1995) investigate the type of people that are excluded from formal finance systems in

Britain and point out that a necessary financial service for low-income people is a basic every bank account [2].

Naceur and Ghazounai (2007) recognized that underdeveloped financial systems has a negative impact on economic growth by studying the relationship between financial **development** and economic growth for 11 regional countries of Middle East and North Africa [3]. Naceur and Ghazounai (2007) also showed that the underdeveloped financial system in the Middle East and North Africa region slowed down economic growth in the region.

There is a close correlation between Human Development Index (HDI) and the Index of Financial Inclusion (IFI). Sarma and Pais (2011) used regression method to investigate the relation between HDI and IFI. The authors suggest that socio-economic and infrastructure related factors, income and physical infrastructure for connectivity and information are important factors of financial inclusion[4].

Besides, Pradhan et al. (2016) analyzed the relationship between insurance market penetration and economic growth. This study investigates the causal interaction of insurance market penetration, broad money, stock market capitalization and economic growth focusing on the ASEAN (Association of South East Asian Nations) ARF (Regional Forum). This study proves that a short-term two-way relationship exists between the insurance market and economic growth [5].

Xuan Vinh Vo et al., (2016) investigates the role of financial structure in promoting economic growth in Vietnam. The empirical result of this study presented that there is no strong relationship between stock market development and economic growth. The impulse response tests and the variance decomposition analysis present evidence of the one-way impact from the stock market capitalization variable on the change of economic growth, however, this effect is marginal [6].

Kim & Hassan (2018) examines the role of financial inclusion on economic growth by using panel data of 55 member countries of the Organization of Islamic Cooperation (OIC) during the 1990 – 2013 period. In order to provide convincing empirical evidence, the authors apply various quantitative methods, such as GMM, VAR, and Granger causality analysis. The results of the analysis show that financial inclusion contributes to the economic growth of these countries [7]. Also examining the relationship between these two variables at a macro level, Makina & Walle (2019) focuses on the case of some African countries, where people's financial access is low [8]. In particular, the research study not only focuses on the impact of financial

inclusion individually but also considers this effect with the interaction of financial development of each country. The results of the analysis once again confirm the role of financial inclusion in economic growth, but the results are inconsistent between different models. Estimated coefficient of financial inclusion is positive but it is not statistical significance when interaction term of financial inclusion and financial development is added to the model.

Recently, Hong Van et al., (2019) analyzed the relationship between financial inclusion and economic growth, especially in emerging countries. The results support a positive relationship between financial inclusion and economic growth. Countries with low income and a lower degree of financial inclusion has stronger relationship between financial inclusion and economic growth. From this results, Policy makers should implement financial inclusion to promote economic growth [9].

From the above researches, we can see that there is not much researches about different of financial inclusion between emerging countries and developed countries. This study focuses on the impact of financial inclusion on economic growth and compares the different impacts between emerging countries and developed countries. Similarly, Sharma (2016) shows a positive effect of financial inclusion on economic growth in the case of India[10]. In this research, the author measures financial inclusion access under a variety of indicators and perspectives, including the penetration of the banking system, the level of access to financial services of people, and results of using banking services.

### III. DATA AND REGRESSION MODEL

We collect data from the Global Financial Development Database published by the World Bank which support the data for over 200 countries. But there are some countries that have no data of all variables used in this research. Therefore, after excluding countries that have missing values, we choose 58 countries to analyze the relationship between financial inclusion and economic growth. The length of data is 12 years from 2006 to 2017. Before the year 2006, there are many countries that the data are not available. Following Sarma (2008), we construct indexes of financial inclusion (IFI) for countries from three aspects, namely accessibility; availability and the actual usage of financial services. Accessibility is measured by bank account per 1000 adults and account at a formal financial institution. Availability is measured by ATMs per 100,000 adults and bank branches per 100,000 adults. Actual usage is measured by credit services and deposit services.

Similar Sarma (2008), the dimension index for the  $i^{\text{th}}$  dimension,  $d_i$ , is computed by the following formula.

$$d_i = \frac{A_i - m_i}{M_i - m_i} \quad (1)$$

Where:

$A_i$  = actual value of dimension  $i$

$m_i$  = minimum value of dimension  $i$

$M_i$  = maximum value of dimension  $i$

Equation (1) ensures that  $0 \leq d_i \leq 1$ . Higher the value of  $d_i$ , higher the country's achievement in dimension  $i$ . The ideal point of  $d_i$ , actual value is equal maximum value, is equal to 1. From calculated  $d_i$ , the index of financial inclusion is computed as follows:

$$IFI_i = 1 - \frac{\sqrt{(1-d_1)^2 + (1-d_2)^2 + \dots + (1-d_n)^2}}{\sqrt{n}} \quad (2)$$

Based on the data collected from World Bank, we calculated IFI. The higher value of  $d_i$  of a country is, the higher value of IFI is. From calculated results, Spain has the highest value of IFI which is 0.642, next is Luxembourg and Portugal. Egypt has the lowest value of IFI which is 0.025. Vietnam has the value of 0.085 and rank 57 in the table.

TABLE I. AVERAGE RESULTS OF IFI FOR THE PERIOD 2006-2017

No	Economy	IFI	No	Economy	IFI
1	Spain	0.642	30	Greece	0.304
2	Luxembourg	0.605	31	Iran	0.303
3	Portugal	0.552	32	Turkey	0.282
4	United States	0.550	33	Chile	0.282
5	Canada	0.542	34	Finland	0.276
6	Australia	0.523	35	Brazil	0.276
7	South Korea	0.478	36	Romania	0.256
8	France	0.472	37	UAE	0.241
9	Belgium	0.452	38	Malaysia	0.240
10	Switzerland	0.444	39	Hungary	0.232
11	England	0.443	40	Bosnia	0.221
12	Croatia	0.426	41	Saudi Arabia	0.213
13	Japan	0.426	42	Uruguay	0.211
14	Israel	0.420	43	Qatar	0.208
15	Italy	0.416	44	Kazakhstan	0.202
16	Bulgaria	0.415	45	Guatemala	0.197
17	Slovenia	0.405	46	Colombia	0.195
18	Ireland	0.401	47	Panama	0.194
19	Estonia	0.375	48	South Africa	0.191
20	Latvia	0.367	49	Ecuador	0.172
21	Germany	0.366	50	Argentina	0.159
22	Austria	0.357	51	Peru	0.142
23	Thailand	0.346	52	China	0.140
24	Sweden	0.335	53	Indonesia	0.135
25	Poland	0.331	54	Mexico	0.135
26	Slovakia	0.324	55	India	0.128
27	Russia	0.317	56	Jamaica	0.114
28	Norway	0.313	57	Vietnam	0.084
29	Netherlands	0.313	58	Egypt	0.025

Source: Calculated by authors

Besides IFI, other variables are also collected from Global Financial Development Database. That are HCAP, TRADE, POLITICAL, PRO\_RIGHT and GOV\_E.



Descriptive statistics of all used variables are presented in table 2.

TABLE II. DESCRIPTIVE STATISTICS

Variable	Obs.	Mean	Std. Dev	Min	Max
IFI	689	0.313	0.155	0.003	0.783
HCAP	606	56.963	23.087	9.315	136.603
TRADE	696	88.982	55.045	22.106	408.362
POLITICAL	692	0.210	0.809	-2.009	1.512
PRO_RIGHT	570	4.685	1.011	2.426	6.606
GOV_E	571	3.792	0.796	2.117	5.663
GDP	696	25674	23456	1079	111968

Source: Calculated by authors

From table 2, HCAP is Human Capital measured by the gross percentage of tertiary enrollments over population for each country. TRADE is the trade openness measured by the percent of import and export to GDP. POLITICAL is political stability and absence of violence. PRO\_RIGHT is Intellectual property right index. GOV\_E is government effectiveness. GDP is Gross Domestic Product. For GDP, we take the logarithm of GDP before estimate regression equation. All above variables are collected from World Bank.

From above data, we analyze the relationship between IFI and logarithm of GDP. We draw a scatter diagram and found a positive relationship between financial inclusion and economic growth. This relationship is presented in figure 1.

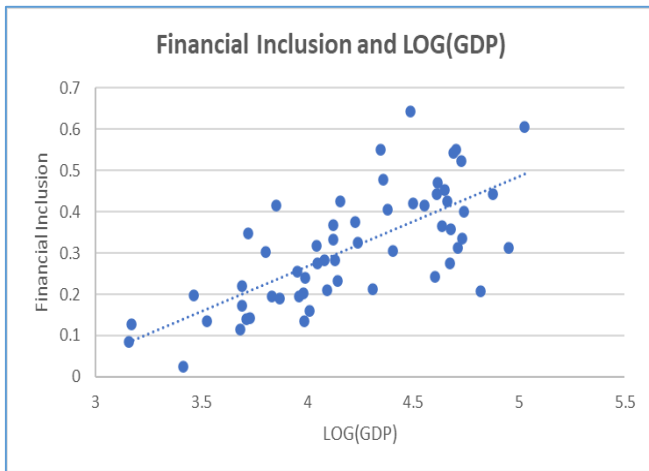


Fig. 1. Relationship between financial inclusion and economic growth

To quantify the relationship between financial inclusion and economic growth, we estimate the dynamic regression equation which presented in equation 3.

$$GDP_{i,t} = \beta_0 + \beta_1 * GDP_{i,t-1} + \beta_2 * IFI_{i,t} + \beta_3 * EMG_{i,t} + \beta_4 * IFI * EMG_{i,t} + \beta_5 * X_{i,t} + \varepsilon_{i,t} \quad (3)$$

$\beta_0$  to  $\beta_5$  is coefficients of the regression.  $GDP_{i,t-1}$  is the lag of  $GDP_{i,t}$ .  $X_{i,t}$  is control variables of country  $i$  at year  $t$ .  $X_{i,t}$  includes HCAP, TRADE, POLITICAL, PRO\_RIGHT.  $\varepsilon_{i,t}$  is the error term of the model. In order to estimate equation 3, we apply GMM (Generalized Method of Moment) because

we use the lag of GDP as independent variable which cause endogenous problem. GMM method can overcome endogenous problem.

#### IV. EMPIRICAL RESULTS

We used Stata software and the command XTABOND2 to estimate the equation 3. In order to check the consistent of the regression model, we analyse three models which vary in independent variables. Every equation has 491 observations and 54 groups. We also test of AR and Sargant test of overidentify. All the P\_value of AR(1) test are less than 5% and the P\_value of AR(2) test are greater than 5%. These mean that the error term have no serial correlation. P\_value of Sargan tests are greater than 5% means that there are not over identification in regression model. All the test results are presented at the end of the table 3.

Based on the analysis results, we found that IFI has positive effect on GPP per capita in three regression models. The significance of IFI variable is less than 1% for all three equation. These results are similar with some previous studies such as: Naceur and Ghazounai (2007), Kim & Hassan (2018), Hong Van et al., (2019) which support the positive relationship between financial inclusion and economic growth. Difference from previous studies, this paper analyze the interaction of financial inclusion index (IFI) and emerging countries (EMG) to check the difference impact of financial inclusion on economic growth.

We also found the impact of EMG and the interaction of IMG and IFI on the economic growth. The significance is less than 1% in equation 1. Based on the regression coefficients, we can say that in emerging countries the impact of financial inclusion on economic growth has smaller slop than developed countries or financial inclusion has more effect on economic growth in developed countries than in emerging countries.

TABLE III. REGRESSION RESULTS BY GMM

Variables	(1) LOG (GDP)	(2) LOG (GDP)	(3) LOG (GDP)
D.LOG (GDP)	0.123 [0.66]	0.180 [0.94]	
EMG	-1.106*** [-12.52]	-1.302*** [-16.87]	-1.294*** [-16.80]
IFI	1.405*** [14.03]	1.154*** [14.08]	1.154*** [14.04]
IFI*EMG	-0.939*** [-4.11]		
Property Right	0.443*** [13.72]	0.474*** [14.56]	0.468*** [14.58]
Human Capital	0.0130*** [3.94]	0.0142*** [4.20]	0.0140*** [4.13]
Trade Openness	0.000468* [1.71]	0.000638** [2.28]	0.000714*** [2.66]
GOV_E	-0.133 [-4.92]	-0.159 [-5.87]	-0.153 [-5.79]
Constant	8.269*** [96.57]	8.298*** [93.98]	8.294*** [93.73]
No of Obs	491	491	491
No of Group	54	54	54
P-value test AR(1)	0.048	0.030	0.029
P-value test AR(2)	0.258	0.129	0.089
P-value Sargan test	0.932	0.053	0.140

Note:  $t$  statistics in brackets; \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Besides, we also found the positive effect of other variables such as Intellectual Property Right, Human Capital and Trade Openness on economic growth. Property Right,

Human Capital has the significance of 1% and Trade openness has the significance of 1% to 10%.

## V. CONCLUSION

We have examined the impact of financial inclusion on economic growth 58 countries during the period 2006 – 2017. The GMM method was used to analyze panel data which collected from World Bank. The analysis results show that the financial inclusion has a positive effect on economic growth in developed and emerging countries. The impacts of financial inclusion on economic growth in developed countries has a steeper slope than in emerging countries. Besides, trade openness and intellectual property right also affect the economic growth of these countries.

From analysis results, to promote economic growth, we can improve financial inclusion by improve accessibility, availability, and the actual usage of financial services. Besides, we can increase import, export, and intellectual property right to promote economic growth. This research has some findings but there are still some limitations that need to do further research. Firstly, we can estimate the impacts of financial inclusion on economic growth by other method like Panel Vector Autoregression or Fixed Effect Model to compare the results. Secondly, we can collect more data in order to improve the reliability of the estimation. Lastly, we can estimate details of every dimension of financial inclusion to find the results in more details.

## REFERENCES

- [1] A. Leyshon and N. Thrift, "The restructuring of the U.K. financial services industry in the 1990s: a reversal of fortune?," *J. Rural Stud.*, vol. 9, no. 3, pp. 223–241, Jul. 1993, doi: 10.1016/0743-0167(93)90068-U.
- [2] A. Leyshon and N. Thrift, "Geographies of Financial Exclusion: Financial Abandonment in Britain and the United States," *Trans. Inst. Br. Geogr.*, vol. 20, no. 3, p. 312, 1995, doi: 10.2307/622654.
- [3] S. Ben Naceur and S. Ghazouani, "Stock markets, banks, and economic growth: Empirical evidence from the MENA region," *Res. Int. Bus. Financ.*, vol. 21, no. 2, pp. 297–315, Jun. 2007, doi: 10.1016/j.ribaf.2006.05.002.
- [4] M. Sarma and J. Pais, "Financial Inclusion and Development," *J. Int. Dev.*, vol. 23, no. 5, pp. 613–628, Jul. 2011, doi: 10.1002/jid.1698.
- [5] R. P. Pradhan, B. M. Arvin, N. R. Norman, M. Nair, and J. H. Hall, "Insurance penetration and economic growth nexus: Cross-country evidence from ASEAN," *Res. Int. Bus. Financ.*, vol. 36, pp. 447–458, Jan. 2016, doi: 10.1016/j.ribaf.2015.09.036.
- [6] X. V. Vo, H. H. Nguyen, and K. D. Pham, "Financial structure and economic growth: the case of Vietnam," *Eurasian Bus. Rev.*, vol. 6, no. 2, pp. 141–154, 2016, doi: 10.1007/s40821-016-0042-8.
- [7] D. W. Kim, J. S. Yu, and M. K. Hassan, "Financial inclusion and economic growth in OIC countries," *Res. Int. Bus. Financ.*, vol. 43, pp. 1–14, 2018, doi: 10.1016/j.ribaf.2017.07.178.
- [8] D. Makina and Y. M. Walle, *Financial Inclusion and Economic Growth: Evidence From a Panel of Selected African Countries*. Elsevier Inc., 2019.
- [9] L. T. H. Van, A. T. Vo, N. T. Nguyen, and D. H. Vo, "Financial Inclusion and Economic Growth: An International Evidence," *Emerg. Mark. Financ. Trade*, vol. 00, no. 00, pp. 1–25, 2019, doi: 10.1080/1540496X.2019.1697672.
- [10] D. Sharma, "Nexus between financial inclusion and economic growth: Evidence from the emerging Indian economy," *J. Financ. Econ. Policy*, vol. 8, no. 1, pp. 13–36, 2016, doi: 10.1108/JFEP-01-2015-0004.

# Ultimate Bond Strength of Steel Bar Embedded in Sea Sand Concrete under Different Curing Environments

Quoc Khanh Tran  
Faculty of Science and Technology  
Hong Bang International University  
Ho Chi Minh City, Vietnam  
gtk2007.qk@gmail.com

Tri Thuong Ngo  
Faculty of Civil Engineering  
Thuyloi University  
Ha Noi, Vietnam  
trithuong@tlu.edu.vn

Duy Liem Nguyen  
Faculty of Civil Engineering and Applied Mechanics  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
liemnd@hcmute.edu.vn

Ngoc Thanh Tran  
Faculty of Civil Engineering  
Ho Chi Minh City University of Transport  
Ho Chi Minh City, Vietnam  
ngocthanh.tran@ut.edu.vn

**Abstract**— This paper investigated ultimate bond strength of steel bar embedded in sea sand concrete under different curing environments. Total 24 test series of cylinder specimens D150 x 300 mm with embedded steel bar were experienced under pull-out test. Two types of sand, normal sand and sea sand, were evaluated. All specimens were cured in normal water and sea water for 7, 14, 28, 90, 180, and 365 days. The results showed that the failure mode of all specimens was pullout mode and no corrosion was observed on the surface of steel bar embedded in sea sand concrete after 365 days curing in seawater. The ultimate bond strength developed rapidly at an early stage but slowly at later stage of curing. There were no significant differences in the ultimate bond strength between specimen cured in normal water and sea water. The full replacement of normal sand with sea sand had no significant effect on the ultimate bond strength between steel bar and concrete.

**Keywords**— Bond strength, steel bar, sea sand, sea water, pullout test

## I. INTRODUCTION

The demand for aggregate used in concrete has been recently increasing due to the rapid development of infrastructure in developing country. It is reported that the demand for aggregates including coarse and fine aggregate increases at a rate of 5.2% per year and achieves 51.79 billion tons in 2019 [1]. Specifically, most of aggregates are manufactured from natural resources and only 10% aggregates production come from other sources. The large amount of aggregate production obviously causes negative impacts regarding to the natural resource and environment. The noticeable environmental impacts of aggregate production are the landslide, water flow change and flood [2]. Thus, it is important to find new alternative sources for normal aggregate in concrete to adapt our present and future demand.

One of promising alternative sources to replace fine aggregate in concrete, is sea sand because this resource is plentiful and easily exploited without causing environment effect [3]. However, the sea sand has different chemical compositions as compared to normal sand. The chloride ion, one of the most common chemical compositions in sea sand, can impact on the hydration process of cement, accelerate the corrosion of steel bar and further effect on the bond strength

between steel bar and concrete [4]. Thus, the effect of adding sea sand on the steel bar corrosion and bond strength should be clearly understood before applying sea sand concrete to infrastructure.

Unfortunately, there is very little information concerning the effect of adding sea sand on the steel bar corrosion and bond behavior, most of researchers have focused on the mechanical resistance of sea sand concrete. Many researches have reported that the sea sand concrete produced higher compressive strength at early age than ordinary concrete [5-7]. In contrast, some researchers concluded that sea sand concrete exhibited lower compressive strength than ordinary concrete [8-9]. According to the best knowledge of author, only Jau et al. examined corrosion of steel bar in sea sand concrete and no obvious corrosion was observed on the surface of steel bar [10]. Although there are some researchers examined the corrosion level and bond behavior between steel bar and sea sand concrete, the corrosion of steel bar was accelerated by using the electric current technical and consequently the effect of adding sea sand was not clear [11-12].

Recently, Tran et al. [13] investigated the effects of adding sea sand on the compressive strength of concrete under different curing environments. The sea sand was exploited at Phu Quoc island, Kien Giang province, Vietnam. The results showed that the sea sand concrete produced from 8% to 17% higher compressive strength than normal concrete and the full replacement of normal sand with sea sand exhibited the best performance. In addition, the compressive strength of sea sand concrete cured in normal water was higher than that in sea water. However, the effect of sea sand on the steel bar corrosion and bond behavior is still under question.

In order to fill knowledge gaps, this study aims to evaluate the ultimate bond strength of steel embedded in sea sand concrete under different curing environments. The detail objectives are: 1) to evaluate effect of curing time and curing environment on the bond strength, 2) to investigate the effect of adding sea sand on the bond strength.

## II. EXPERIMENTAL PROGRAM

### A. Materials and specimen preparation

An experimental program was set up to evaluate ultimate bond strength of steel bar embedded in sea sand concrete under various curing environments, as shown in Fig.1. Total 24 test series were prepared with at least three specimens per series. Two types of sand, normal sand and sea sand, were investigated. The specimens were cured in two different environments, normal water and sea water, until testing date of 7, 14, 28, 90, 180, and 365 days.

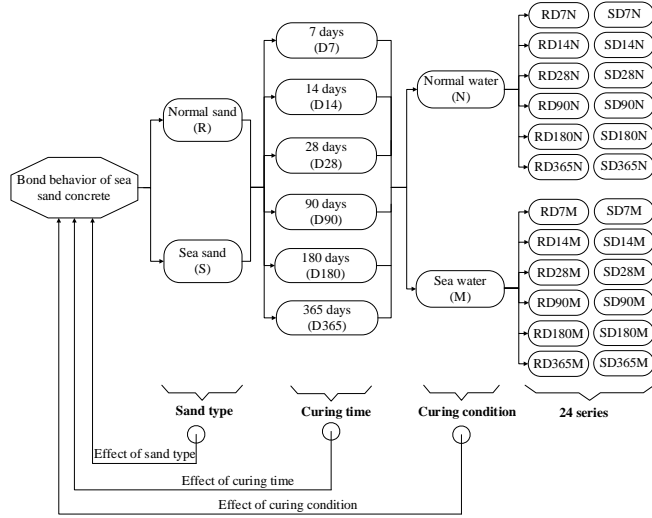


Fig. 1. Detail of experimental program

The composition of mixture and the 28-day compressive strength of 150 mm cubic specimen are given in Table 1. In addition, the images of normal and sea sand are shown in Fig. 2 and their physical properties and chemical composition are provided in Table 2. The sea sand was exploited at Phu Quoc island, Kien Giang province, Vietnam and it was dried for 20 day at room temperature before mixing. The chloride content in sea sand is much higher than that in normal sand (Table 2). On the other hand, the properties of steel bar are summarized in Table 3.

TABLE I. MIXTURE PROPORTIONS OF GEOPOLYMER CONCRETE

Cement (kg/m <sup>3</sup> )	Coarse aggregate (kg/m <sup>3</sup> )	Fine aggregate (kg/m <sup>3</sup> )	Water (kg/m <sup>3</sup> )	Compressive strength (MPa)
465	1150	650	185	35

TABLE II. PHYSICAL PROPERTIES AND CHEMICAL COMPOSITION OF SAND

Material	Specific gravity (g/cm <sup>3</sup> )	Particle size distribution (mm)	Chloride (%)
Normal sand	2.62	0.14 – 5 mm	0.002
Sea sand	2.6	0.14 – 5 mm	0.21

TABLE III. PROPERTIES OF STEEL BAR

Diameter (mm)	Yield strength (MPa)	Ultimate strength (MPa)	Ultimate strain (%)	Surface shape
14	640	740	20	Ribbed



Fig. 2. Images of sand

In the preparation of material, the steel bars with a length of 400 mm was prepared. The length of embedded segment is 200 mm while that of free segment is 200 mm. The embedded length is chosen to assess impacts of sea sand at interfacial zone while maintaining pullout mode under pullout test. The free segment was coated by epoxy and then cover by plastic layer to avoid corrosion. All the steel bars were held tight inside cylinder molds right before mixing. The mixture was mixed by using A Hobart-type mixer with 20-L capacity. Firstly, the coarse aggregate and fine aggregate were put into mixer and dry mixed for 5 min. Then, the cement was put and further mixed for 5 min. After dry mixing, the water was divided two part and put two times and mixed for 5 min. Next, a wide scoop was used to pour the mixture into metal cylinder molds with embedded steel bar and then the fresh mixture was compacted by hand. After 48 h stored in laboratory at room temperature, all the specimens were demolded. After demolding, twelve series were put in normal water and the other twelve series were cured in sea water, as shown in Fig. 3. Artificial sea water was generated by mixing sodium chloride powder (99%) and water with the ratio of 3.5% by weight. The specimens were carried out to perform pullout test at the age of 7, 14, 28, 90, 180 and 365 days.



a) Normal water immersion b) Sea water immersion

Fig. 3. Curing conditions

### B. Test setup and procedure

The test set-ups are illustrated in Fig. 4. The specimen geometry was cylinder with the dimension size of D150×300 mm. The diameter of steel bar was 14 mm and the embedded length was 200 mm. The cylinder specimen with embedded steel bar was held by a steel frame during pullout test. A universal testing machine (UTM) with capacity of 1000 KN was carried out to perform pullout test. The pullout load was recorded by the load cell. The testing procedure was followed to standard TCVN 197-1:2014 [14].

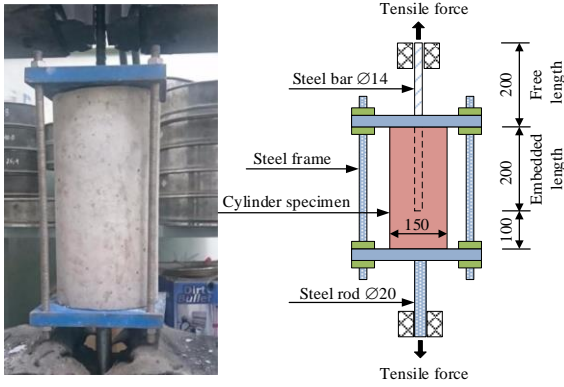


Fig. 4. Pullout test set up

### III. RESULTS AND DISCUSSION

#### A. Ultimate bond strength between steel bar and sea sand concrete

The ultimate bond strength between steel bar and concrete was determined using formula (1):

$$\tau_{\max} = \frac{P_{\max}}{\pi \phi L_{em}} \quad (1)$$

Where  $\tau_{\max}$  is defined as ultimate bond strength,  $P_{\max}$  is defined as maximum pullout load,  $\phi$  is defined as nominal steel bar diameter and  $L_{em}$  is defined as embedded length.

The average ultimate bond strength of all series is summarized in Table 4. The results showed that the ultimate bond strength of steel bar embedded sea sand concrete varied from 5.93 MPa to 8.42 MPa depending on the curing time, curing environment and sand type. In addition, the failure mode of all specimens was pullout mode and no corrosion was observed on the surface of steel bar embedded in sea sand concrete after 365 days curing in sea water, as shown in Fig. 5. In addition, the maximum tensile stress in the steel bar (480 MPa) was smaller than yield strength (640 MPa).

TABLE IV. ULTIMATE BOND STRENGTH BETWEEN STEEL BAR AND CONCRETE

Group 1		Group 2	
Series	Ultimate bond strength (MPa)	Series	Ultimate bond strength (MPa)
RD7N	5.94 (0.10)	SD7N	6.06 (0.09)
RD14N	7.1 (0.10)	SD14N	7.06 (0.16)
RD28N	7.88 (0.12)	SD28N	7.89 (0.09)
RD90N	8.13 (0.08)	SD90N	8.10 (0.09)
RD180N	8.11 (0.17)	SD180N	8.19 (0.12)
RD365N	8.38 (0.07)	SD365N	8.31 (0.12)
Group 3		Group 4	
Series	Ultimate bond strength (MPa)	Series	Ultimate bond strength (MPa)
RD7M	6.08 (0.15)	SD7M	5.93 (0.11)
RD14M	7.12 (0.13)	SD14M	6.74 (0.11)
RD28M	8.11 (0.12)	SD28M	7.42 (0.12)
RD90M	8.23 (0.17)	SD90M	8.11 (0.17)
RD180M	8.27 (0.07)	SD180M	8.38 (0.17)
RD365M	8.42 (0.12)	SD365M	8.38 (0.17)

\*Note: the number in bracket is the standard deviation

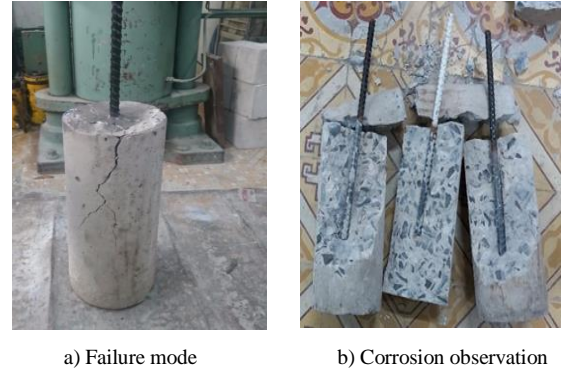


Fig. 5. Failure of pullout specimens after testing at the age of 365 days

#### B. Development of ultimate bond strength between steel bar and sea sand concrete

Fig. 6 shows the effects of curing time on the ultimate bond strength of steel bar in concrete. As the curing time increased, the ultimate bond strength increased. The ultimate bond strength developed rapidly at an early stage but slowly at later stage of curing. In detail, the ultimate bond strength increased 30 - 33% with increasing curing time from 7 days to 28 days but increased 4 - 13% as the curing time increased from 28 days to 365 days. This is due to the fact that the ultimate bond strength between steel bar and concrete was strongly dependent on the compressive strength of concrete and consequently the development trends of the ultimate bond strength were similar to that of the compressive strength [15].

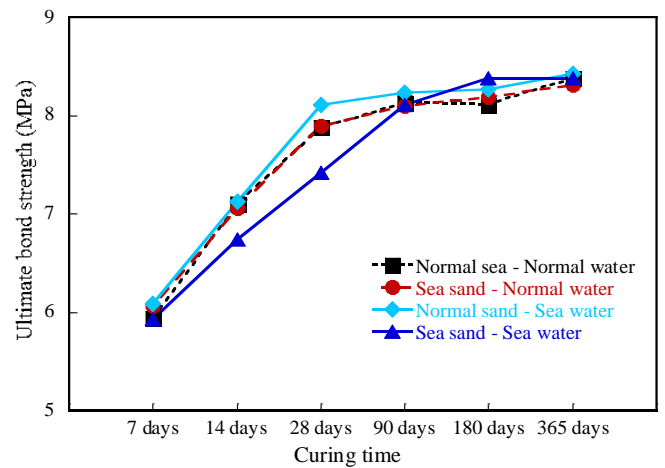


Fig. 6. Effect of curing time on the ultimate bond strength of steel bar in concrete

#### C. Effect of curing environments on the ultimate bond strength between steel bar and sea sand concrete

The effects of curing environments on the ultimate bond strength are shown in Fig. 7. There were no significant differences in the ultimate bond strength between specimen cured in normal water and sea water. The specimens cured in normal water showed 0 - 8% difference in the ultimate bond strength compared to those cured in sea water. It is clear that the chloride content in sea water might not impact on the hydration process and structure development in concrete until 365 days curing. In addition, the chloride content in sea water could not attack and cause corrosion of steel bar due to the sufficient thickness of concrete cover until 365 days.



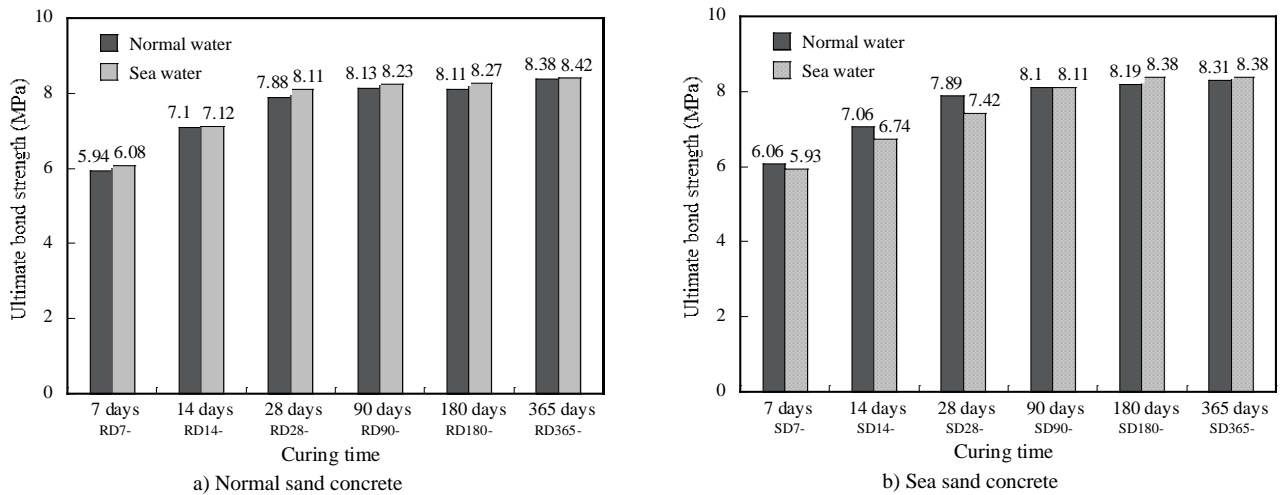


Fig. 7. Effect of curing environments on the ultimate bond strength of steel bar in concrete

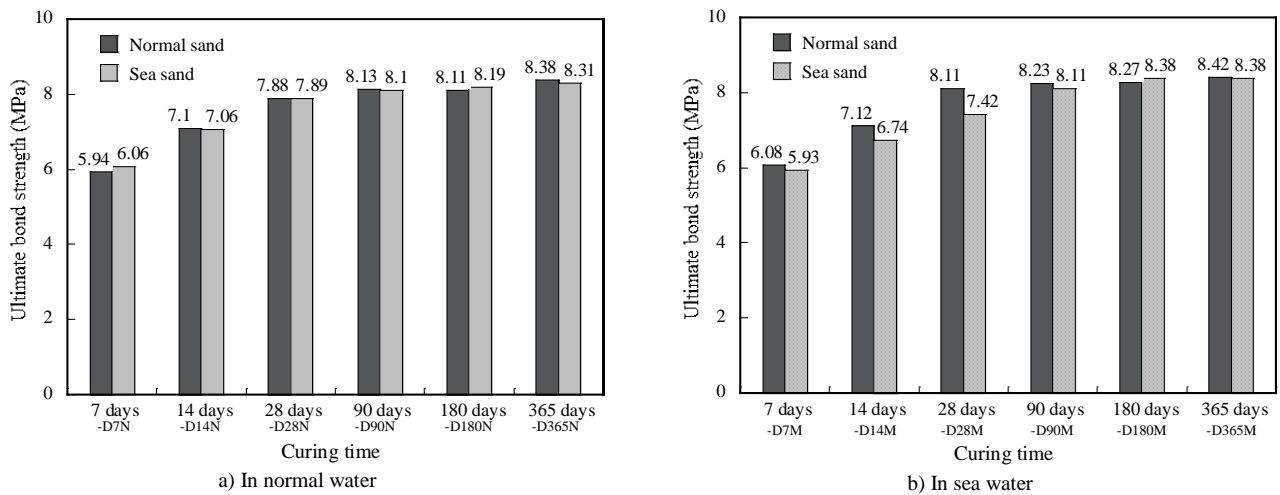


Fig. 8. Effect of adding sea sand on the ultimate bond strength of steel bar in concrete

#### D. Effect of adding sea sand on the ultimate bond strength between steel bar and sea sand concrete

Fig. 8 shows the effects of sea sand on the ultimate bond strength between steel bar and concrete. The full replacement of normal sand with sea sand had no significant effect on the ultimate bond strength between steel bar and concrete. The sea sand concrete resulted in 0 – 3% change in the ultimate bond strength compared to ordinary concrete. Thus, the full replacement of normal sand with sea sand might not have much effect on the hydration process and structure development in concrete and could not activate the corrosion process of steel bar until 365 days curing.

#### IV. CONCLUSION

An experimental program was performed to determine ultimate bond strength of steel bar embedded in sea sand concrete under different curing environments. Based on the results of this study, the following conclusions can be withdrawn:

- The failure mode of all specimens was pullout mode and no corrosion was observed on the surface of steel bar embedded in sea sand concrete after 365 days curing in seawater.

- The ultimate bond strength developed rapidly at an early stage but slowly at later stage of curing. The ultimate bond strength increased 30 – 33% with increasing curing time from 7 days to 28 days but increased 4 – 13% as the curing time increased from 28 days to 365 days.
- There were no significant differences in the ultimate bond strength between specimen cured in normal water and sea water. The specimens cured in normal water showed 0 – 8% difference in the ultimate bond strength compared to those cured in sea water.
- The full replacement of normal sand with sea sand had no significant effect on the ultimate bond strength between steel bar and concrete. The sea sand concrete resulted in 0 – 3% change in the ultimate bond strength compared to ordinary concrete.

#### REFERENCES

- [1] Z. Abdollahnejad, M. Mastali, M. Falah, T. Luukkonen, M. Mazari, and M. Illikainen, "Review Construction and Demolition Waste as Recycled Aggregates in Alkali-Activated Concretes," *Materials*, vol. 12, pp. 1-22, 2019.



- [2] M.D. Gavriltea, "Review Environmental Impacts of Sand Exploitation. Analysis of Sand Market," *Sustainability*, vol. 9(1118), pp. 1-26, 2017.
- [3] Y. Fan, R.H. Zhang, and Z.B. Chen, "Research on the Properties of Sea Sand Mortar," *Applied Mechanics and Materials*, vol. 405, pp. 2871-2875, 2013.
- [4] J. Xiao, C. Qiang, A. Nanni, and K. Zhang, "Use of sea-sand and seawater in concrete construction: Current status and future opportunities," *Construction and Building Materials*, vol. 155, pp. 1101-1111, 2017.
- [5] J. Limeir, L. Agullo, and M. Etxeberria, "Dredged marine sand as construction material," *Euro. J. Environ. Civ. Eng.*, vol. 16, pp. 1-13, 2012.
- [6] S.R. Chandrakerthy, "Suitability of sea sand as a fine aggregate for concrete production," *Trans. Inst. Eng.*, pp. 93-114, 1994.
- [7] S.D. Ramaswamy, M.A. Aziz, and C.K. Murthy, "Sea dredged sand for concrete, extending aggregate resources," *ASTM Int.*, vol. 774, pp. 167-177, 1982.
- [8] C.G. Girish, D. Tensing, and K.L. Priya, "Dredged offshore sand as a replacement for fine aggregate in concrete," *Int. J. Eng. Sci. Emerg. Technol.*, vol. 8, pp. 88-95, 2015.
- [9] I.H. Cagatay, "Experimental evaluation of buildings damaged in recent earthquakes in Turkey," *Eng. Fail. Anal.*, vol. 12(3), pp. 440-452, 2005.
- [10] W.C. Jau, J.C. Tan, and C.T. Yang, "Effect of sea sand on concrete durability and its management," *J Southeast Univ. Nat. Sci. Ed.*, vol. 36, pp. 160-166, 2006.
- [11] W.P.S. Dias, G.A.P.S.N. Seneviratne, and S.M.A. Nanayakkara, "Offshore sand for reinforced concrete," *Constr. Build. Mater.*, vol. 22(7), pp. 1377-1384, 2008.
- [12] L.B. Bian, S.M. Song, and F. Li, "Experimental study on durability of sea sand concrete," *China Concr. Cem. Prod.*, vol. 2, pp. 11-14, 2012.
- [13] N.T. Tran, N.H. Nguyen, M.T. Duong, and T.D. Le, "Evaluation of compressive strength of concrete using sea sand under various curing environment," *Journal of Science and Technology in Civil Engineering*, vol. 14, pp. 60-72, 2020).
- [14] TCVN 197-1:2014, "Metallic materials - Tensile testing - Part 1: Method of test at room temperature," 2014.
- [15] A.M. Diab, H.E. Elyamany, M.A. Hussein, and H.M. Al Ashy, "Bond behavior and assessment of design ultimate bond stress of normal and high strength concrete," *Alexandria Engineering Journal*, vol. 53, pp. 355-371, 2014.

# Performance Analysis and Evaluation of Underlay Two-Way Cooperative Networks with NOMA

Nguyen Duc Anh

Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
anhnguyenduc883@gmail.com

Pham Ngoc Son

Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
sonpndtvt@hcmute.edu.vn

**Abstract**—In this paper, we investigate two-way decode-and-forward scheme with non-orthogonal multiple access (NOMA) in underlay cognitive networks to increase efficiency of spectrum utilization as well as requirements of future networks. Secondary users communicate with each other via a secondary relay to decode and forward signal using the NOMA technology. We analyze and evaluate system performance in terms of outage probabilities with Rayleigh fading channels. The theoretical analyses are exactly validated by the Monte Carlo simulation method. The analysis and simulation results show that the system performance is improved by higher interference constraint thresholds and by closer locations to the secondary users of the cooperative relay.

**Keywords**— *Non-orthogonal multiple access (NOMA), Cognitive network, Two-way cooperation, Outage probability.*

## I. INTRODUCTION

Compared with orthogonal multiple access (OMA), non-orthogonal multiple access (NOMA) technique is used to solve the problem of improving the channel capacity in 5G networks [1-3]. These issues have attracted the attention of many researchers as well as related topics. In [2], the authors studied about performance of NOMA technology in the power domain and compared with the OMA technique applied in previous generation mobile communication networks. The results have shown that the NOMA technique is more effective than the OMA. In addition, the authors in [3] developed the NOMA technique in full-duplex two-way relaying communications. However, the authors only compared the effectiveness of the NOMA technique with the OMA without analyzing and evaluating this technique in cognitive networks.

The cognitive network has attracted increasing interests to solve the problem of exhaustion of frequency spectrum [4-6]. In [4], the authors proposed and evaluated the cognitive network model combining NOMA technology and compared with OMA technique. The results have shown that the system performance is improved when the NOMA technique is applied in cognitive networks. However, the authors did not consider and evaluate the system performance with two-way communication which also offers desirable benefits such as: speed and spectrum efficiency.

Inspired by the above ideas, in this paper, we investigate underlay two-way cooperative networks with using the NOMA. The distribution of this paper is mathematical analysis, evaluation of factors affecting the underlay cognitive network that combines both two-way communication and NOMA technique at the cooperative

relay. This paper desires to solve the challenges of 5G network as well as apply to sensor networks, ad-hoc networks to increase the efficiency of frequency spectrum sharing and performance.

The main contributions of the paper are summarized as follows. Firstly, we present an underlay two-way cooperative network with the NOMA. Secondly, we analyze the system performance via the outage probabilities. Thirdly, the performance system is improved by increasing the interference constraint thresholds and by closer locations to the secondary users of the cooperative relay. Finally, the outage probabilities with Rayleigh fading channels are derived and are confirmed by Monte Carlo simulations.

This paper is organized as follows: Section II describes a system model of the underlay two-way cooperative networks with the NOMA; Section III analyzes and calculates the outage probability; the simulation results are presented in Section IV; and Section V summarizes our conclusions.

## II. SYSTEM MODEL

The underlay two-way cooperative network with the NOMA is described in Figure 1. This two-way relaying model operates the half-duplex communication with two time slots in which a relay R performs the successive interference cancellation (SIC) to decode the desired data. At the first time slot, two secondary sources S1 and S2 transmit data  $x_1$  and  $x_2$  to the secondary relay, respectively. At the second time slot, the relay R decodes  $x_1$  and  $x_2$ , and creates a coded data  $x$  by the digital network coding (XOR operation) as  $x = x_1 \oplus x_2$ . Next, the relay R transmits the data  $x$  to the S1 and S2. Operations in two time slots need to consider interference constraints in a primary user Pr. The Pr in the Fig.1 can be an eavesdropper to take the data  $x_1$  and  $x_2$  illegally [7-9]. The authors in [7-11] investigated the physical layer security protocols to against the eavesdropper by relay selections aided artificial noise, jammer, energy harvesting, and opportunistic operations in two-way networks.

In Fig.1,  $(h_1, d_1)$ ,  $(h_2, d_2)$ ,  $(h_3, d_3)$ ,  $(h_4, d_4)$ ,  $(h_5, d_5)$  are fading channel coefficients and the normalized link distances of links S1-R, S2-R, S1-Pr, S2-Pr and R-Pr, respectively. In the case that nodes S1, S2, R and Pr are moving, the channel coefficients  $h_i$  will be modelled as double Rayleigh fading [12],  $i \in \{1, 2, 3, 4, 5\}$ .

In this paper, we set some assumptions reasonably as follows: each node has a private antenna, variances of zero-mean White Gaussian Noises (AWGNs) are identical,

denoted similarly as  $N_0$ . In addition, all channels  $h_i$  are considered in Rayleigh fading channels [10, 11]. Hence, the channel gains are  $g_i = |h_i|^2$  which have exponential distributions with parameters  $\lambda_i = d_i^\beta$ , where  $\beta$  is a path-loss exponent,  $i \in \{1, 2, 3, 4, 5\}$  [13].

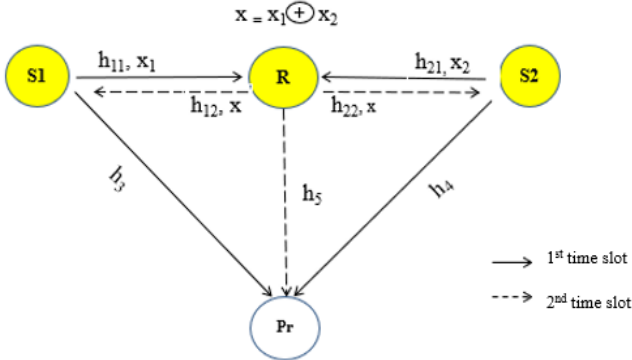


Fig. 1. System model of two-way underlay cognitive network with the NOMA

### III. PERFORMANCE ANALYSIS

For simpler analyses, estimated channel state information (CSI) and transceiver hardware of wireless devices are considered perfectly. Analyses of the system performance in the imperfect CSIs and hardware impairments have been discussed clearly in [6, 14, 15]. Under interference constraint threshold  $I$  of the primary receiver Pr, the transmit powers of S1, S2 and R are set maximally as  $P_{S1} = \alpha I / g_3$ ,  $P_{S2} = (1 - \alpha) I / g_4$ ,  $P_R = I / g_5$  respectively, where  $\alpha$  is power allocation coefficient exchanging between S1 and S2 ( $0 < \alpha < 1$ ) [6, 16].

At the first time slot, the received signal at R is expressed by:

$$y_R = \sqrt{P_{S1}} h_{11} x_1 + \sqrt{P_{S2}} h_{21} x_2 + n_R \quad (1)$$

where the second subscript 1 of channel coefficients  $h_{11}$  and  $h_{21}$  indicates the first time slot.

\*If  $d_1 \leq d_2$ : By using the SIC, firstly the relay R decodes  $x_1$  before decoding  $x_2$ . Hence, signal-to-interference-plus-noise ratio (SINR) at the relay R for decoding  $x_1$  is given by:

$$\gamma_{R \rightarrow x_1} = \frac{P_{S1} g_{11}}{P_{S2} g_{21} + N_0} = \frac{(\alpha I / g_3) g_{11}}{((1 - \alpha) I / g_4) g_{21} + N_0} = \frac{\alpha Q X}{(1 - \alpha) Q Y + 1} \quad (2)$$

where:  $Q = \frac{I}{N_0}$ ,  $X = \frac{g_{11}}{g_3}$  and  $Y = \frac{g_{21}}{g_4}$ .

After decoding  $x_1$ , the signal at the R after cancelling the known component  $\sqrt{P_{S1}} h_{11} x_1$  is obtained as follow:

$$y_{R-x_1} = \sqrt{P_{S2}} h_{21} x_2 + n_R \quad (3)$$

Next, the relay R decodes  $x_2$ . The SINR at R for decoding  $x_2$  is given from (3) as follow:

$$\begin{aligned} \gamma_{R \rightarrow x_2} &= \frac{P_{S2} g_{21}}{N_0} = \frac{((1 - \alpha) I / g_4) g_{21}}{N_0} \\ &= (1 - \alpha) Q \frac{g_{21}}{g_4} = (1 - \alpha) Q Y \end{aligned} \quad (4)$$

At the second time slot, the relay R transmits  $x = x_1 \oplus x_2$  to the secondary sources Sk,  $k \in \{1, 2\}$ . The received signal at the Sk is presented as follow:

$$y_{Sk} = \sqrt{P_R} h_{k2} x + n_{Sk} \quad (5)$$

where the second subscript 2 of channel coefficients  $h_{k2}$  indicates the second time slot.

The SINR at the Sk for decoding x is given by:

$$\gamma_{Sk \rightarrow x} = \frac{P_R g_{k2}}{N_0} = \frac{(I / g_5) g_{k2}}{N_0} = Q \frac{g_{k2}}{g_5} \quad (6)$$

Achievable data rates at the R and the Sk are given as follow:

$$R_{R \rightarrow x_k} = \frac{1}{2} \log_2 (1 + \gamma_{R \rightarrow x_k}) \quad (7)$$

$$R_{Sk \rightarrow x} = \frac{1}{2} \log_2 (1 + \gamma_{Sk \rightarrow x}) \quad (8)$$

where the fraction 1/2 indicates that the system operates in the two time slots.

The outage probability at S2 to decode  $x_1$  is expressed by a math expression:

$$P_{S2} = \Pr \left[ \underbrace{R_{R \rightarrow x_1} < R_{Th}}_{\Phi_1} \right] + \Pr \left[ \underbrace{R_{R \rightarrow x_1} \geq R_{Th}, R_{S2 \rightarrow x} < R_{Th}}_{\Phi_2} \right] \quad (9)$$

where  $R_{Th}$  is a target data rate, bits/s/Hz.

In (9), the probability  $\Phi_1$  is manipulated as

$$\begin{aligned} \Phi_1 &= \Pr \left[ R_{R \rightarrow x_1} < R_{Th} \right] = \Pr \left[ \frac{1}{2} \log_2 (1 + \gamma_{R \rightarrow x_1}) < R_{Th} \right] \\ &= \Pr \left[ \gamma_{R \rightarrow x_1} < 2^{2R_{Th}} - 1 \right] \\ &= \Pr \left[ \frac{\alpha Q X}{(1 - \alpha) Q Y + 1} < 2^{2R_{Th}} - 1 \right] \\ &= \Pr \left[ \alpha Q X < (2^{2R_{Th}} - 1) ((1 - \alpha) Q Y + 1) \right] \\ &= \Pr \left[ X < \frac{1 - \alpha}{\alpha} (2^{2R_{Th}} - 1) Y + \frac{2^{2R_{Th}} - 1}{\alpha Q} \right] \\ &= \Pr \left[ X < \tau_1 Y + \tau_2 \right] \end{aligned} \quad (10)$$

where  $\tau_1 = \frac{1 - \alpha}{\alpha} (2^{2R_{Th}} - 1)$  and  $\tau_2 = \frac{2^{2R_{Th}} - 1}{\alpha Q}$ .

The cumulative distribution function (CDF) of the random variable X is obtained as follow:

$$\begin{aligned}
 F_X(y) &= \Pr[X < y] = \Pr\left[\frac{g_{11}}{g_3} < y\right] \\
 &= \int_0^\infty F_{g_{11}}(yz) \times f_{g_3}(z) dz \\
 &= \int_0^\infty (1 - e^{-\lambda_1 yz}) \lambda_3 e^{-\lambda_3 z} dz \\
 &= 1 - \frac{\lambda_3}{\lambda_3 + \lambda_1 y} = \frac{\lambda_1 y}{\lambda_3 + \lambda_1 y}
 \end{aligned} \quad (11)$$

Similarly, the CDF of the random variable Y is solved as follow:

$$F_Y(y) = \Pr[Y < y] = \Pr\left[\frac{g_{21}}{g_4} < y\right] = \frac{\lambda_2 y}{\lambda_4 + \lambda_2 y} \quad (12)$$

The probability density function (PDF) of the random variable Y is inferred as

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y} = \frac{\lambda_2 \lambda_4}{(\lambda_4 + \lambda_2 y)^2} \quad (13)$$

By using (11) and (13), we have a closed-form expression for  $\Phi_1$  in (10) as follow:

$$\begin{aligned}
 \Phi_1 &= \int_0^\infty f_Y(y) \times F_X(\tau_1 y + \tau_2) dy \\
 &= \int_0^\infty \left[ \frac{\lambda_2 \lambda_4}{(\lambda_4 + \lambda_2 y)^2} \left( 1 - \frac{\lambda_3}{\lambda_3 + \lambda_1 (\tau_1 y + \tau_2)} \right) \right] dy \\
 &= 1 - \int_0^\infty \frac{\lambda_2 \lambda_3 \lambda_4}{(\lambda_4 + \lambda_2 y)^2 (\lambda_3 + \lambda_1 (\tau_1 y + \tau_2))} dy \\
 &= 1 - \frac{\lambda_2 \lambda_3 \lambda_4}{\lambda_1 \tau_1 \tau_3} \left( \frac{1}{\lambda_4} + \frac{1}{\tau_3} \ln \left( \frac{\lambda_4}{\lambda_4 + \tau_3} \right) \right)
 \end{aligned} \quad (14)$$

$$\text{where } \tau_3 = \frac{\lambda_2 \lambda_3 + \lambda_1 \lambda_2 \tau_2 - \lambda_1 \lambda_4 \tau_1}{\lambda_1 \tau_1}.$$

We note that the last integral in (14) is solved by performing variable transformation as  $t = \lambda_4 + \lambda_2 y$ .

In (9), the probability  $\Phi_2$  is presented and manipulated as

$$\begin{aligned}
 \Phi_2 &= \Pr[R_{R \rightarrow x_1} \geq R_{Th}, R_{S2 \rightarrow x} < R_{Th}] \\
 &= \underbrace{\Pr[R_{R \rightarrow x_1} \geq R_{Th}]}_{\Phi_{21}} \times \underbrace{\Pr[R_{S2 \rightarrow x} < R_{Th}]}_{\Phi_{22}}
 \end{aligned} \quad (15)$$

where  $\Phi_{21}$  and  $\Phi_{22}$  are solved as

$$\Phi_{21} = \Pr[R_{R \rightarrow x_1} \geq R_{Th}] = 1 - \Pr[R_{R \rightarrow x_1} < R_{Th}] = 1 - \Phi_1 \quad (16)$$

$$\begin{aligned}
 \Phi_{22} &= \Pr[R_{S2 \rightarrow x} < R_{Th}] = \Pr[\gamma_{S2 \rightarrow x} < 2^{2R_{Th}} - 1] \\
 &= \Pr\left[Q \frac{g_{22}}{g_5} < 2^{2R_{Th}} - 1\right] \\
 &= \Pr\left[\frac{g_{22}}{g_5} < \frac{2^{2R_{Th}} - 1}{Q}\right] = \frac{\lambda_2 \tau_4}{\lambda_5 + \lambda_2 \tau_4}
 \end{aligned} \quad (17)$$

$$\text{where } \tau_4 = \frac{2^{2R_{Th}} - 1}{Q}.$$

Substituting (16) and (17) into (15), we have an expression of  $\Phi_2$  as follow:

$$\Phi_2 = \Phi_{21} \times \Phi_{22} = (1 - \Phi_1) \times \frac{\lambda_2 \tau_4}{\lambda_5 + \lambda_2 \tau_4} \quad (18)$$

Substituting (14) and (18) into (9), the outage probability at S2 to decode  $x_1$  is obtained as follow:

$$P_{S2} = 1 - \frac{\lambda_2 \lambda_3 \lambda_4 \lambda_5}{\lambda_1 \tau_1 \tau_3 (\lambda_5 + \lambda_2 \tau_4)} \left( \frac{1}{\lambda_4} + \frac{1}{\tau_3} \ln \left( \frac{\lambda_4}{\lambda_4 + \tau_3} \right) \right) \quad (19)$$

The outage probability at S1 to decode  $x_2$  is obtained as follow:

$$\begin{aligned}
 P_{S1} &= \Pr\left[ \underbrace{R_{R \rightarrow x_1} < R_{Th}}_{\Phi_1} \right] + \Pr\left[ \underbrace{R_{R \rightarrow x_1} \geq R_{Th}, R_{R \rightarrow x_2} < R_{Th}}_{\Phi_3} \right] \\
 &\quad + \Pr\left[ \underbrace{R_{R \rightarrow x_1} \geq R_{Th}, R_{R \rightarrow x_2} \geq R_{Th}, R_{S1 \rightarrow x} < R_{Th}}_{\Phi_4} \right]
 \end{aligned} \quad (20)$$

In which, the probability  $\Phi_3$  is expressed and can be manipulated as

$$\begin{aligned}
 \Phi_3 &= \Pr[R_{R \rightarrow x_1} \geq R_{Th}, R_{R \rightarrow x_2} < R_{Th}] \\
 &\approx \Pr[R_{R \rightarrow x_1} \geq R_{Th}] \times \Pr[R_{R \rightarrow x_2} < R_{Th}] \\
 &= (1 - \Pr[R_{R \rightarrow x_1} < R_{Th}]) \times \Pr[R_{R \rightarrow x_2} < R_{Th}] \\
 &= (1 - \Phi_1) \times \Pr[R_{R \rightarrow x_2} < R_{Th}]
 \end{aligned} \quad (21)$$

In (21),  $\Pr[R_{R \rightarrow x_2} < R_{Th}]$  is solved as

$$\begin{aligned}
 \Pr[R_{R \rightarrow x_2} < R_{Th}] &= \Pr[\gamma_{R \rightarrow x_2} < 2^{2R_{Th}} - 1] \\
 &= \Pr\left[(1 - \alpha) Q \frac{g_{21}}{g_4} < 2^{2R_{Th}} - 1\right] \\
 &= \Pr\left[\frac{g_{21}}{g_4} < \tau_5\right] = \frac{\lambda_2 \tau_5}{\lambda_4 + \lambda_2 \tau_5}
 \end{aligned} \quad (22)$$

$$\text{where } \tau_5 = \frac{2^{2R_{Th}} - 1}{(1 - \alpha) Q}.$$

Substituting (22) into (21), we can obtain an expression of  $\Phi_3$  as follow:

$$\Phi_3 = (1 - \Phi_1) \times \frac{\lambda_2 \tau_5}{\lambda_4 + \lambda_2 \tau_5} \quad (23)$$

The probability  $\Phi_4$  in (20) can be solved similarly as

$$\begin{aligned} \Phi_4 &= \Pr[R_{R \rightarrow x_1} \geq R_{Th}, R_{R \rightarrow x_2} \geq R_{Th}, R_{S1 \rightarrow x} < R_{Th}] \\ &\approx \frac{\lambda_1 \lambda_4 \tau_4 (1 - \Phi_1)}{(\lambda_4 + \lambda_2 \tau_5)(\lambda_5 + \lambda_1 \tau_4)} \end{aligned} \quad (24)$$

Substituting (14), (23) and (24) into (20), the outage probability at S1 to decode  $x_2$  is obtained as follow:

$$\begin{aligned} P_{S1} &\approx \Phi_1 + (1 - \Phi_1) \times \frac{\lambda_2 \tau_5}{\lambda_4 + \lambda_2 \tau_5} + \frac{\lambda_1 \lambda_4 \tau_4 (1 - \Phi_1)}{(\lambda_4 + \lambda_2 \tau_5)(\lambda_5 + \lambda_1 \tau_4)} \\ &= \Phi_1 + \frac{(1 - \Phi_1)}{(\lambda_4 + \lambda_2 \tau_5)} \left( \lambda_2 \tau_5 + \frac{\lambda_1 \lambda_4 \tau_4}{\lambda_5 + \lambda_1 \tau_4} \right) \end{aligned} \quad (25)$$

\*If  $d_1 > d_2$ : Also, by using the SIC, firstly the relay R decodes  $x_2$  before decoding  $x_1$ .

Because of the symmetry system model as shown in Fig.1, the outage probability at S1 to decode  $x_2$  is presented and then is inferred from (9) and (19) as

$$\begin{aligned} P_{S1} &= \Pr[R_{R \rightarrow x_2} < R_{Th}] + \Pr[R_{R \rightarrow x_2} \geq R_{Th}, R_{S1 \rightarrow x} < R_{Th}] \\ &= 1 - \frac{\lambda_1 \lambda_3 \lambda_4 \lambda_5}{\lambda_2 \tau_1 \tau_6 (\lambda_5 + \lambda_1 \tau_4)} \left( \frac{1}{\lambda_3} + \frac{1}{\tau_6} \ln \left( \frac{\lambda_3}{\lambda_3 + \tau_6} \right) \right) \end{aligned} \quad (26)$$

$$\text{where } \tau_6 = \frac{\lambda_1 \lambda_4 + \lambda_1 \lambda_2 \tau_2 - \lambda_2 \lambda_3 \tau_1}{\lambda_2 \tau_1}$$

Similarly, the outage probability at S2 to decode  $x_1$  is expressed and obtained as

$$\begin{aligned} P_{S2} &= \Pr[R_{R \rightarrow x_2} < R_{Th}] + \Pr[R_{R \rightarrow x_2} \geq R_{Th}, R_{R \rightarrow x_1} < R_{Th}] \\ &\quad + \Pr[R_{R \rightarrow x_2} \geq R_{Th}, R_{R \rightarrow x_1} \geq R_{Th}, R_{S2 \rightarrow x} < R_{Th}] \\ &\approx \Phi_5 + \frac{(1 - \Phi_5)}{(\lambda_3 + \lambda_1 \tau_5)} \left( \lambda_1 \tau_5 + \frac{\lambda_2 \lambda_3 \tau_4}{\lambda_5 + \lambda_2 \tau_4} \right) \end{aligned} \quad (27)$$

$$\text{where } \Phi_5 = 1 - \frac{\lambda_1 \lambda_3 \lambda_4}{\lambda_2 \tau_1 \tau_6} \left( \frac{1}{\lambda_3} + \frac{1}{\tau_6} \ln \left( \frac{\lambda_3}{\lambda_3 + \tau_6} \right) \right).$$

#### IV. SIMULATION RESULTS

In this section, the system performance is analyzed and evaluated using theoretical analyses and Monte Carlo simulations of outage probabilities. In two-dimensional plane, the coordinates of nodes are S1(0,0), S2(1,0), Pr(0.5, -1) and R( $x_R$ ,  $y_R$ ), where  $0 < x_R < 1$ . Hence, the link distances S1-R, R-S2, S1-Pr, S2-Pr and R-Pr, are obtained as

follow  $d_1 = \sqrt{x_R^2 + y_R^2}$ ,  $d_2 = \sqrt{(1 - x_R)^2 + y_R^2}$ ,  $d_3 = d_4 = 1.12$  and  $d_5 = \sqrt{(0.5 - x_R)^2 + (1 + y_R)^2}$ , respectively. It is assumed that the path-loss exponent is set to a constant as  $\beta = 3$  and  $Q$  (dB) on the x-axis is defined as  $Q = 10 \times \log_{10}(I/N_0)$  (dB). Markers imply simulated values and solid lines denote analyzed results.

Figure 2 presents the outage probability at the S1 and S2 versus  $Q$  (dB) when  $R_{Th} = 0.5$  (bit/s/Hz),  $R(x_R=0.1, y_R=0)$  and  $\alpha = 0.5$ . Hence,  $d_1 = 0.1$  and  $d_2 = 0.9$  ( $d_1 < d_2$ ). In this case ( $d_1 < d_2$ ), the outage probabilities at S1 and S2 decrease when the  $Q$  (dB) increases in an interval  $Q$  (dB)  $\in [0, 35]$  and then reach the saturation values ( $Q > 35$  (dB)). These can be explained that when the  $Q$  increases, the SINRs  $\gamma_{R \rightarrow x_1}$ ,  $\gamma_{R \rightarrow x_2}$  and  $\gamma_{S1 \rightarrow x}$  will increase. Therefore, the achievable data rates  $R_{R \rightarrow x_1}$  and  $R_{S1 \rightarrow x}$  increase. It means that the ability at the S1 and S2 to decode successfully is improved. Hence, the system performance depends on interference constraint levels of the primary network. In addition, we can see that the simulation results fit well to the theoretical ones. It proves the acceptability of theoretical expressions in the section III.

Figure 3 presents the outage probabilities at the S1 and S2 versus  $Q$  (dB) when  $R_{Th} = 0.5$  (bit/s/Hz),  $R(x_R=0.9, y_R=0)$  and  $\alpha = 0.5$ . Hence,  $d_1 = 0.9$  and  $d_2 = 0.1$  ( $d_1 > d_2$ ). As shown in Fig. 3, the outage probabilities at the S1 and S2 decrease when the  $Q$  (dB) increases and also reach the saturation levels ( $Q > 35$  (dB)).

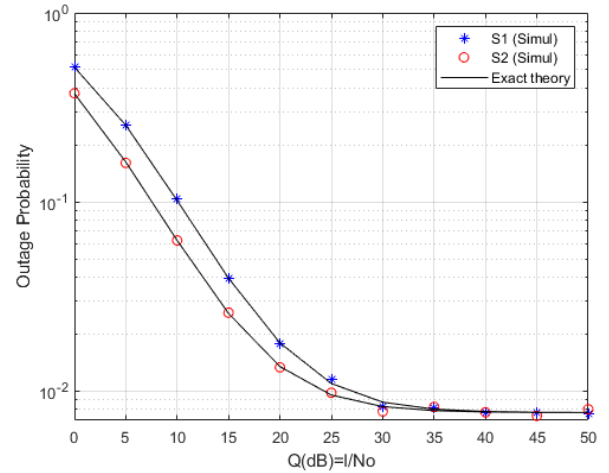


Fig. 2. The outage probabilities at S1 and S2 versus  $Q$  (dB) ( $d_1=0.1, d_2=0.9$ )



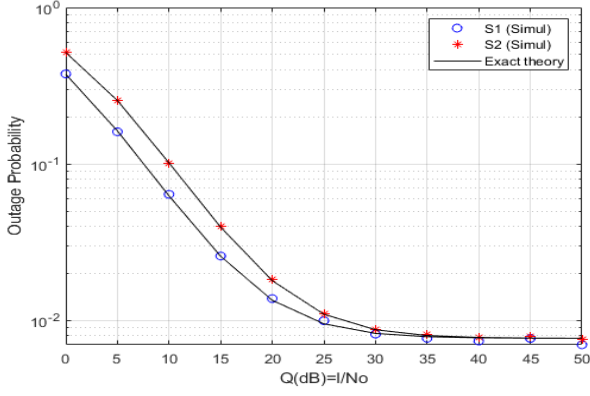
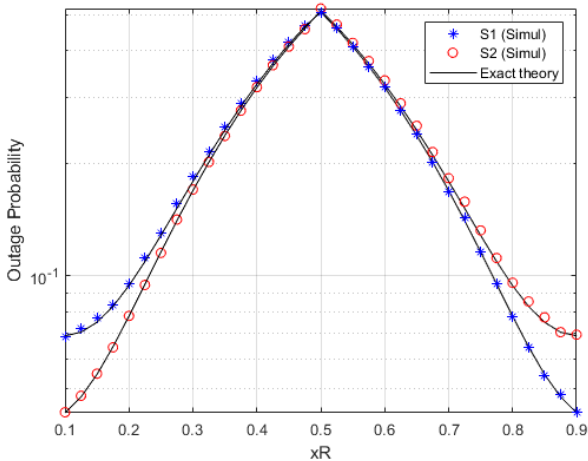
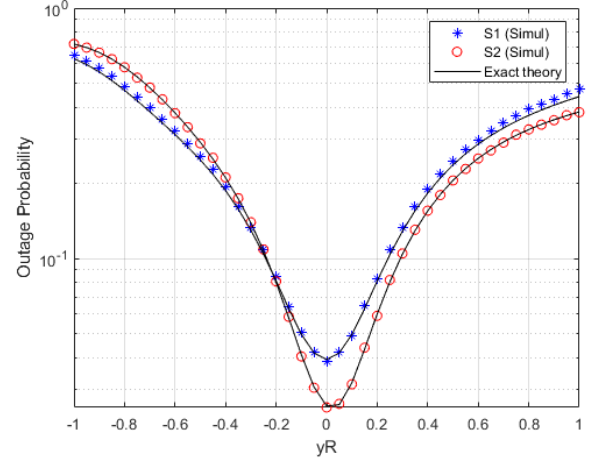

 Fig. 3. The outage probabilities at S1 and S2 versus  $Q$  (dB) ( $d_1=0.9, d_2=0.1$ )

Figure 4 presents the outage probabilities at S1 and S2 versus  $x_R$  ( $0.1 \leq x_R \leq 0.9$ ) when  $R_{th} = 0.5$  (bit/s/Hz),  $Q = 12$  (dB) and  $\alpha = 0.5$ . Hence,  $d_1 = x_R$  and  $d_1 = 1 - x_R$  with both two cases  $d_1 > d_2$  and  $d_1 \leq d_2$ . As shown in Fig. 4, if  $x_R$  increases when  $x_R \in (0.1, 0.5)$ , the outage probabilities increase and the outage probabilities at the S2 are smaller than at the S1. If  $x_R$  increases when  $x_R \in (0.5, 0.9]$ , the outage probabilities occur in the opposite situations. In addition, the performances achieve the best values when R is located at  $x_R=0.1$  and  $x_R=0.9$ . However, at the midpoint of S1 and S2, the system performances reach the worst values. Therefore, the system performances change and depend on the location of the cooperative relay R.

Figure 5 presents the outage probabilities at S1 and S2 versus  $y_R$  when  $R_{th} = 0.5$  (bit/s/Hz),  $Q = 12$  (dB) and  $\alpha = 0.5$ . Hence,  $d_1 = \sqrt{0.1^2 + y_R^2}$  and  $d_2 = \sqrt{0.9^2 + y_R^2}$ . From the Fig 5, the outage probabilities at the S1 and S2 change versus  $y_R$  and achieve the lowest level at  $y_R=0$ . In addition, we confirm that the simulation results match well to the theoretical values.


 Fig. 4. The outage probabilities at S1 and S2 versus  $x_R$ 

 Fig. 5. The outage probability at the S1 and S2 versus  $y_R$ 

## V. CONCLUSIONS

In this paper, we investigated the underlay two-way cooperative networks with using the NOMA. We analyzed and evaluated the system performance via the outage probabilities. In addition, the outage probabilities with Rayleigh fading channels were derived and confirmed by Monte Carlo simulations. The investigations showed that the system performance is improved by increasing the interference constraint thresholds and by closer locations to the secondary users of the cooperative relay.

## ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2019.13.

## REFERENCES

- [1] X. Zou, B. He, and H. Jafarkhani, "An Analysis of Two-User Uplink Asynchronous Non-orthogonal Multiple Access Systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1404-1418, 2019.
- [2] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721-742, 2017.
- [3] X. Wang, M. Jia, I. W. Ho, Q. Guo, and F. C. M. Lau, "Exploiting Full-Duplex Two-Way Relay Cooperative Non-Orthogonal Multiple Access," *IEEE Transactions on Communications*, vol. 67, no. 4, pp. 2716-2729, 2019.
- [4] L. Lv, J. Chen, Q. Ni, Z. Ding, and H. Jiang, "Cognitive Non-Orthogonal Multiple Access with Cooperative Relaying: A New Wireless Frontier for 5G Spectrum Sharing," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 188-195, 2018.
- [5] Z. Zhang, Y. Lu, Y. Huang, and P. Zhang, "Neural Network-Based Relay Selection in Two-Way SWIPT-Enabled Cognitive Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6264-6274, 2020.
- [6] P. N. Son, "Joint impacts of hardware impairments, imperfect CSIs, and interference constraints on underlay cooperative cognitive networks with reactive relay selection," *Telecommunication Systems*, vol. 71, no. 1, pp. 65-76, 2019.
- [7] S. Q. Nguyen and H. Y. Kong, "Improving Secrecy Outage and Throughput Performance in Two-Way Energy-Constraint Relaying Networks Under Physical Layer Security," *Wireless*

- Personal Communications*, vol. 96, no. 4, pp. 6425-6457, 2017.
- [8] P. N. Son and H. Y. Kong, "Exact Outage Probability of Two-Way Decode-and-Forward Scheme with Opportunistic Relay Selection Under Physical Layer Security," *Wireless Personal Communications*, vol. 77, no. 4, pp. 2889-2917, 2014.
  - [9] P. N. Son and H. Y. Kong, "An Integration of Source and Jammer for a Decode-and-Forward Two-way Scheme Under Physical Layer Security," *Wireless Personal Communications*, vol. 79, no. 3, pp. 1741-1764, 2014.
  - [10] Z. Cao, X. Ji, J. Wang, S. Zhang, Y. Ji, and J. Wang, "Security-Reliability Tradeoff Analysis for Underlay Cognitive Two-Way Relay Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 6030-6042, 2019.
  - [11] X. Ding, T. Song, Y. Zou, X. Chen, and L. Hanzo, "Security-Reliability Tradeoff Analysis of Artificial Noise Aided Two-Way Opportunistic Relay Selection," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 3930-3941, 2017.
  - [12] T. T. Duy, G. C. Alexandropoulos, V. T. Tung, V. N. Son, and T. Q. Duong, "Outage performance of cognitive cooperative networks with relay selection over double-Rayleigh fading channels," *IET Communications*, vol. 10, no. 1, pp. 57-64, 2016.
  - [13] P. N. Son and H. Y. Kong, "An approach of Relay ordering to improve OFDM-based cooperation," *IEICE Transactions on Communications*, vol. E98-B, no. 5, pp. 870-877, 2015.
  - [14] X. Sun, K. Xu, and Y. Xu, "Performance analysis of multi-pair two-way amplify-and-forward relaying with imperfect CSI over Ricean fading channels," *IET Communications*, vol. 12, no. 3, pp. 261-270, 2018.
  - [15] B. C. Nguyen, X. N. Tran, D. T. Tran, X. N. Pham, and L. T. Dung, "Impact of Hardware Impairments on the Outage Probability and Ergodic Capacity of One-Way and Two-Way Full-Duplex Relaying Systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8555-8567, 2020.
  - [16] H. Kim, S. Lim, H. Wang, and D. Hong, "Optimal Power Allocation and Outage Analysis for Cognitive Full Duplex Relay Systems," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3754-3765, 2012.

# Reusing Fabric Scraps in Garment Industry - A Green Manufacturing Process

Thi Bich Dung Phung

Faculty of Garment Technology and Fashion Design,  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
dungptb@hcmute.edu.vn

Tuan-Anh Nguyen

Faculty of Garment Technology and Fashion Design,  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
nta@hcmute.edu.vn

**Abstract**—In textile and garment industry, waste disposal management plays an important role to not only increase manufacturing efficiency but also keep working environment in stability as well as economic benefits. Vietnam is the top country in apparel production over the world, in which extremely large textile materials are used to process garment products, a huge amount of fabric scraps is consequently released to the surroundings. This work is to investigate the usability of fabric scraps through an overview of clothing production in Vietnam. Accordingly, authors would like to propose creative production solutions in reducing, reusing, recycling, recovering and landfilling in order to develop useful products in human life such as handicrafts, stuffings and denatured materials

**Keywords**—Textile waste, textile pollution, fabric scrap, garment company, rag recycling

## I. INTRODUCTION

Textiles as well as apparel industry is considered as the largest and fastest growing field owing to increasing population, applications, productivity [1]. It is known to be the world's second pollution resource occupied about 10% of the global carbon dioxide (CO<sub>2</sub>) emission after the oil and petrochemical industry [2]. Such disadvantages in textile production processes are very serious global challenges for managers and scientists to look for the best solutions to decrease the negative effects [3]. Among them, the textile recycling approaches were believed to be the most remarkable solutions. Textile recycling is to recover and reprocess materials into useful products, which is able to reduce significantly negative impacts on life environment [4].



Figure 1. Symbol of textile recycling (fibre2fashion.com)

It is estimated that people annually dispose at least 75% of textile wastes owing to the absences of feasible recycling strategies. Most seriously, more than 85% of end-use clothes went to landfill and burning. Meanwhile, environmentally friendly textile materials are being limited. Particularly, the global cotton growing area is being quickly shrunk due to the desertification as well as urbanization and people must use the pesticides, chemical fertilizers, and defoliant at very high

level to increase the productivity consequently but it causes the largely toxic residues on the end-use products. In addition, several wet dyeing and finishing processes also discharge the hazardous substances and additionally these consumes a huge amount of fossil energy which raises the greenhouse effects. Eventually, the apparel processing and using operations also induced a lot of unwanted compounds such as enzymes, oxidants and dyeing residues. Especially, about 65% of textile fibers are produced by chemical synthesis through polymerization reactions, also called as artificial fibers which induces extremely carbon dioxides to atmosphere [5,6]. Main environmental problem of synthetic textile fibers is not biodegradable. It is evidenced that old clothes from polyester, acrylic, nylon and spandex fabrics take hundreds of years to decompose as they are discharged in landfill. In the world, only 12% of disposed textiles are mechanically recycled by shredding into fibers and other products. Clothes reuse can significantly contribute to reducing the environmental burden. Many challenges as well as difficulties are obviously exposed to minimize the bad effects of textiles on human life.

In clothing industry, the production generates various textile wastes. Fabric scrap accounts for the majority with the highest ratio in the cutting stages as listed in Table I [7]

TABLE I. DISTRIBUTION PERCENTAGE OF FABRIC SCRAPS

Stages	Mainly generated wastes	Percentage
Planning	Paper	0.0%
Warehousing	Fabric scrap, paper, trims	3.0%
Designing	Fabric scrap, paper, plastic	0.5%
Sampling	Fabric scrap, trims	0.5%
Cutting	<b>Fabric scrap</b> , sander paper	<b>92.0%</b>
Sewing	Fabric scrap, trims, needle, oil	4.0%
Finishing	Trims	0.0%
Packing	Paper, plastic, trims, adhesive tape	0.0%



Figure 2. Fabric scraps of (a) small and (b) large pieces in garment company's cutting room before discharged to landfill.

Fabric scraps are known as textile solid wastes which the reduction process is reviewed for a long-term period with requiring appropriate management in the short-term periods avoiding environmental and health problems for humans [8].

Recently, Vietnam's textile and apparel industries are rapidly developing to become a world's leading exporter with the CMT and FOB manufacturing modes which dispose a huge amount of fabric scraps from cutting department. There are some given strategies to solve trashes which may summarized as follows:

- (1) Decrease fabric wastes in cutting department by increasing the highest marker efficiency (i.e, the ratio between pattern pieces area and total marker area)
- (2) Optimize equipment and production to decrease generated toxic substances
- (3) Revoke, classify and manage effectively fabric wastes
- (4) Reuse and recycle fabric scraps to make other products and seek consumer market for them
- (5) Send bio-gradable textile wastes such as cotton and wool to landfills
- (6) Propose strategies to periodically reduce textile wastes for all manufacturing operations
- (7) Restrict uses of hazardous compounds in textile and garment manufacturing
- (8) Build strict awareness for all employees in following the waste managing regulations and labor safety.

According to the international and national publications on textile recycling solutions, it can affirm that although authors have generalized many ways to solve fabric scraps but the obtained results from such studies are not easy to specifically apply for small-size and medium-size enterprises in Vietnam because of many differences in economic scale among countries and in production organization among garment companies. Therefore, in this work, the authors would build some particular solutions associated with handling the solid wastes (i.e, fabric scraps) obtained from cutting operations in garment factories. The aim of the work is to investigate the usability of recycled garments by evaluating fabric scraps.

## II. METHODOLOGY

The method used in the work consists of identifying the literature review through databases that deal with the fabric scraps on domestic and international publishing systems (Scopus, Web of Science and other journals) [1-3, 9]. Especially, based on analyzing the data obtained from the observations and estimations during the internship and field trip periods at the large local garment companies in regards to the marker making and fabric cutting activities. In addition, the experts working in the related fields were interviewed via the available questions to obtain the information about cloth rag as well as the recycling procedures. Accordingly, the authors have suggested some specific solutions to reuse the fabric scraps obtained from the garment companies.

## III. RESULTS AND DISCUSSION

### A. Significances and aims of reusing fabric scraps

Fabric wastes mainly obtained from the following types: (1) defected fabrics in weaving, knitting, dyeing, printing and finishing processes, (2) product samples which are not used after completing the purchasing order, (3) cloth sheets of roll ends which are removed upon the request of garment production, (4) damaged fabrics and clothes in assembling stages, and (5) cloth rags on cutting table. The feasible and right solutions to use such textile pieces can bring a lot of benefits for not only self-enterprises but also social community.

As mentioned previously, fabric scraps from cutting room of garment companies should be reduced down to the lowest level because of more negative impacts on life environment. It has been suggested that the garment enterprises might increase their profits through the savings of input materials and expenses. Besides, these ideas provide not only more jobs but also higher incomes for labors if the fabric rags are fully utilized to create and trade out new products with more conveniences and selections for costumers.



Figure 3. Significances of recycling of fabric scraps

Furthermore, by creating useful items from fabric rags, a great number of jobs need to hire the disadvantaged children or disable persons, thereby the garment enterprises may demonstrate their community responsibilities and charity policies as well.

### B. Solutions of processing and managing fabric scraps

At the holistic level, to manage efficiently the waste in textile industries, the best solutions should be done as:

- (1) Promulgating the specific policies and models for classification of wastes in garment factories to build the labor's environmental preventing awareness
- (2) Monitoring, controlling and penalizing the infringed production activities such direct discharging wastes to the life environment or polluted trash handling
- (3) Encouraging the garment enterprises to use the innovative optimized equipment and eliminating the contaminated waste disposing processes.
- (4) Establishing a network to connect garment companies to handmade workshops, employment centers (for orphans, disable persons), vocational schools, manufacturing or startup companies to reuse rags

Based on the reality, almost garment companies deliver the rags to the local urban environmental services at a certain cost. However, the treatment capacity of these services does not meet the environmental requirements if burying or burning methods are selected normally, because of inducing



very negative impacts on the global climate. Therefore, this work offers five simple and feasible specific recommendations in terms of the classification methods (particularly in the next sections) that a garment company may choose to solve the fabric scraps as follows:

(1) If the fabric scraps are large enough to create other clothes or industrial rags, they should be sold out the affordable markets at cheaper cost level than the original price.

(2) Very small pieces (usually considered as trashes within allowance as well as contract) should be recycled and reused in terms of materials and dimensions.

(3) The classification of wastes or rags needs to be daily regulated according to the 5S model in the cutting room.

(4) Building manufacturing lines in place to produce the largely dimensional fabric scraps as well as increase profits for the whole company.

(5) Offering or selling the small fabric scraps to the business premises (bathroom carpet, coaster, denatured materials) or the job centers for charity to decrease the waste disposal costs or increase the total profit or express the social responsibilities with the local community

#### C. Sorting fabric scraps in terms of original materials

To ensure the classification rags easily, a set of guidance should be promulgated for all cutting staffs. Such materials need to put into types, i.e., interlacing method (woven or knitted fabric) and original material (cotton/synthetic fiber) with two various trash cans per each type as show in **Figure 4**. This sorting method is extremely necessary since it not only saves time but also provides further information for designers to select the most appropriate materials in case of woven and knitted fabrics. By this way, the enterprise easily offers the suitable trash solving directions such as landfill, burning and recycling.

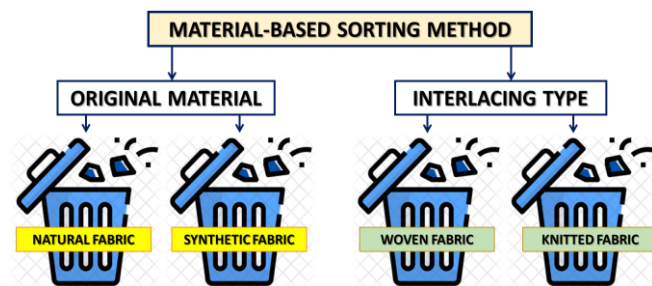


Figure 4. Sorting fabric scraps in terms of material types

As reusing fabric scraps into clothes, both product's technical and aesthetic requirements are given to meet their end-uses. For example, the fabric to design the kid's wears (e.g, face mask, gloves, socks, hat, pillow, blanket, so on) requires to be breathable and absorbent, which is suitable with natural fabrics or blended fabrics with more than 80% of cotton. In other cases, the backpack, handbag, wallet, covered button, and painting cloths are able to use entirely the synthetic or blended fabrics.

To join appropriately several tiny fabric pieces together as sewing blankets or backpack, all patterns should be similar to ensure the shrinkage and dimensional stability for finished products. Therefore, the woven and knitted fabrics have to separate away due to the difference of recovery under stress.

Obviously, the classification of scraps in garment companies requires the combination of policies for cutting staffs, especially in building worker's awareness and working rules.

#### D. Sorting fabric scraps in terms of piece area

Upon cutting operation, the scrapped fabric pieces should be divided into two types according to their sizes (i.e., large or small pieces) in order to determine the specific values for each one. This work proposed a very quick and simple sorting way of fabric rags, namely "hand rule", as described in the following illustration:

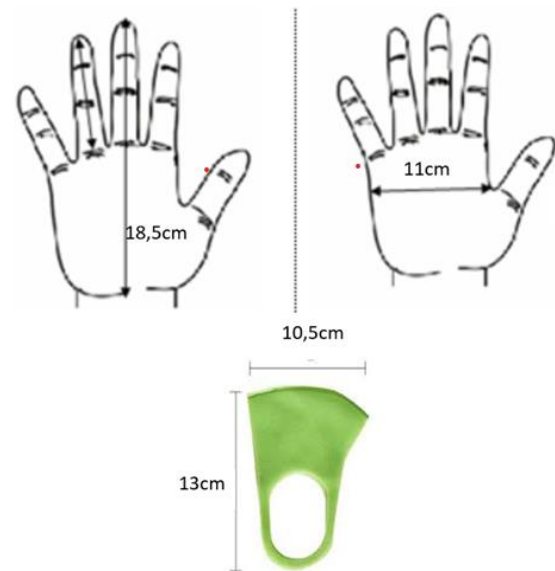


Figure 5. Comparing dimension between a men worker's hand at the age of 25-30 and kid's face mask



Figure 6. Some design ideas from fabric scraps such as mask, purse, artificial flower, coaster and handbag

Based on the anthropometric characteristics of human body, the hand dimensions (**HD**) are used to standardize the sorting steps for scrap pieces (**SP**) at three levels (desired both width and length):

(1)  $SP \geq HD$ : the piece is larger than the hand. These materials can be used to make face mask, children's mitten, newborn's hat/cap, apron, covered button, coaster, cloth flowers, cloth pictures, kid's toys, etc.

(2)  $1/2 HD > SP > HD$ : the piece is between a half and whole hand. They can be created pillow cases, two-layer blankets, carpet, backpack and handcrafted gifts.

(3)  $SP \leq 1/2 HD$ : the piece is smaller than the hand. Because of too tiny sizes, these are not enough to make the new items. Accordingly, they should be supplied for contractors to shred into small particles to mix other powers to produce the heat/electric/sound insulated sheets, wall/floor tiles, hot pot coasters and carpets, etc.

#### E. Applying sorting criteria to create specific products

After the classification criteria for fabric rags are displayed above including interlacing way, original material and piece area according to the “hand rule”, the authors offer the specific applications for each group as shown in Table II.

TABLE II. CLASSIFICATION CRITERIA OF FABRIC SCRAPS

Group	Piece Descriptions	End-uses
A	<ul style="list-style-type: none"> <li>• <math>SP \geq HD</math></li> <li>• Do not join pieces together</li> <li>• Woven or knitted fabrics</li> <li>• 80-100% natural fibers</li> </ul>	Kid’s mask, glove, mitten, newborn hat, apron, covered button, coaster, souvenir’s fabric, key fob, toys, painting fabric
B	<ul style="list-style-type: none"> <li>• <math>1/2 HD &gt; SP &gt; HD</math></li> <li>• Require to join pieces together</li> <li>• Woven or knitted fabrics</li> <li>• 80-100% natural fibers</li> </ul>	Adult’s mask, pillow case, two-layer blanket, kid’s wear, upholstery, seating carpet, coaster, key fob
C	<ul style="list-style-type: none"> <li>• <math>1/2 HD &gt; SP &gt; HD</math></li> <li>• Require to join pieces together</li> <li>• Woven or knitted fabrics</li> <li>• Synthetic fibers</li> </ul>	Cloth flowers, covered button, souvenir’s cover, hair bow, toys, coaster, pot mat, pencil case, wallet or purse, handbag, backpack
D	<ul style="list-style-type: none"> <li>• <math>SP \leq 1/2 HD</math></li> <li>• Require to join pieces together</li> <li>• Woven or knitted fabrics</li> <li>• 80-100% natural fibers</li> </ul>	Bathroom carpet, hot mat, blanket, pillow case, scarf, cushion with quilting, pearl fabric, reused natural textile materials
E	<ul style="list-style-type: none"> <li>• <math>SP \leq 1/2 HD</math></li> <li>• Require to join pieces together</li> <li>• Woven or knitted fabrics</li> <li>• Synthetic fibers</li> </ul>	Covered button, cloth flower, art handcrafted gift with quilting, re materials in techniques such as tile, decoration and insulation

TABLE III. CLASSIFICATION CRITERIA OF FABRIC SCRAPS

Interlacing types		Descriptions	Group
FABRIC SCRAPS	Knitted fabrics	>80% natural fibers $SP > HD$	A
		>80% natural fibers $1/2 HD > SP > HD$	B
		>80% natural fibers $SP < 1/2 HD$	E
		Synthetic fibers $SP > HD$	C
		Synthetic fibers $1/2 HD > SP > HD$	C
		Synthetic fibers $SP < 1/2 HD$	E
	Woven fabric	>80% natural fibers $SP > HD$	A
		>80% natural fibers $1/2 HD > SP > HD$	B
		>80% natural fibers $SP < 1/2 HD$	D
		Synthetic fibers $SP > HD$	C
		Synthetic fibers $1/2 HD > SP > HD$	C
		Synthetic fibers $SP < 1/2 HD$	E



Table II and III suggest specific creative ideas for designers as selecting the various sorted fabric pieces. For example, to make a pillow case, the designer can look for in Table II to choose the product names in group A or D, then according to Table III, the type of fabric exhibits more than 80% of natural fibers and its appropriate area should be between half and full of hands for both knitted and woven fabrics.

#### IV. CONCLUSION

Based on reviewing and investigating the previous studies and the reality, this work analyzes and estimated many negative impacts of fabric scraps on ecological environment and human life. Accordingly, the authors did specifically propose the feasible and simple classification methods of textile wastes (i.e. fabric rags) in terms of material origin and dimensions adapted to the reality in the garment factories to control and decrease bad effect. The given solutions in the paper not only increase the enterprise's significant benefits and profits but also express deeply the social responsibilities with the locals in the sustainable development.

#### ACKNOWLEDGMENT

Authors would like to thank for financial supporting from Ho Chi Minh City University of Technology and Education (HCMUTE). They also express their thanks for experts at garment companies who provided useful information and databases.

#### REFERENCES

- [1] Katherine Le (2018), Textile Recycling Technologies, Colouring and Finishing Methods, UBC Sustainability Scholar
- [2] Athina Koligkioni et al (2018), Environmental Assessment of End-of-Life Textiles in Denmark, *Procedia CIRP*, 69, 962-967, <https://doi.org/10.1016/j.procir.2017.11.090>
- [3] Pensupa N., et al., (2017), Recent trends in sustainable textile waste recycling methods: current situation and future prospects. *Top Curr Chem (Z)*. 2017; 375:76, <https://doi.org/10.1007/s41061-017-0165-0>.
- [4] Hawley, J. M (2006), Textile Recycling: A System Perspective: A System Perspective, *Woodhead Publishing Series in Textiles*, 7-24, <https://doi.org/10.1533/9781845691424.1.7>
- [5] Gustav Sandin, Greg M Peters (2018), Environmental Impact of Textile Reuse and Recycling - A Review, *Journal of Cleaner Production*, 184, 353-365, <https://doi.org/10.1016/j.jclepro.2018.02.266>
- [6] Laura Farrant, Stig Irving Olsen, Arne Wangel (2010), Environmental benefits from reusing clothes, *International Journal of Life Cycle Assessment*, 15, 726-736, <https://doi.org/10.1007/s11367-010-0197-y>
- [7] Hawley J.M (2014), Textile Recycling, *Handbook of Recycling, Textile and Apparel Management*, 1, 211-217, <https://doi.org/10.1016/B978-0-12-396459-5.00015-5>
- [8] Eliane Pinheiro1 and Antonio Carlos de Francisco (2016), Management and Characterization of Textile Solid Waste in a Local Productive Arrangement, *Fibres & Textiles in Eastern Europe*; 24, 4(118): 8-13. <https://doi.org/10.5604/12303666.1201128>
- [9] J Barua-Ramos, et al. (2017), Social and Economic Importance of Textile Reuse and Recycling in Brazil, *Materials Science and Engineering*, 254(19), 192003, <http://doi:10.1088/1757-899X/254/19/192003>

# Analyzing Total Quality Management of Service Enterprises in Vietnam

Nguyen Thi Anh Van  
Department of Industrial Management  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
anhvan@hcmute.edu.vn

Nguyen Khac Hieu  
Department of Industrial Management  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
hieunk@hcmute.edu.vn

**Abstract**—This paper analyzes Total Quality Management (TQM) application in services industries in Vietnam. TQM is measured by seven factors such as: context of the organization, leadership, planning, support, operation, performance evaluation, and improvement. The regression model was used with data collected from 1525 enterprises in nine services sectors. The descriptive statistics shows that the lowest rate of TQM application is 4.29% of the Education and Training sector, while the highest proportion is for the health sector (32.61%). The regression results illustrate that TQM has a positive effect on firm performance( revenue and profit). From the results, some solutions are proposed to enhance the performance of service enterprises in Vietnam through TQM factor improvements.

**Keywords**—organizational innovation, firm performance, manufacturing enterprises.

## I. INTRODUCTION

In the current integrated economy, when Vietnam joins global trade organizations such as WTO and AFTA, the competition among enterprises is getting fiercer. In the past, the government made use of tariff tools or technical barriers to protect the domestic industry, but the integration of those tools was no longer effective, so quality is the most important factor for businesses to improve their competitive position. However, the number of Vietnamese enterprises having QM certificates in accordance with ISO 9001: 2015 is very low. According to the results of a number of surveys of the International Organization for Standardization (ISO), the number of Vietnamese enterprises receiving ISO 9001 (certified for quality management system) is not high. This certification rate is not only much lower than that of developed countries, but also significantly lower than neighboring countries such as Thailand, Malaysia (Figure 1) [1]

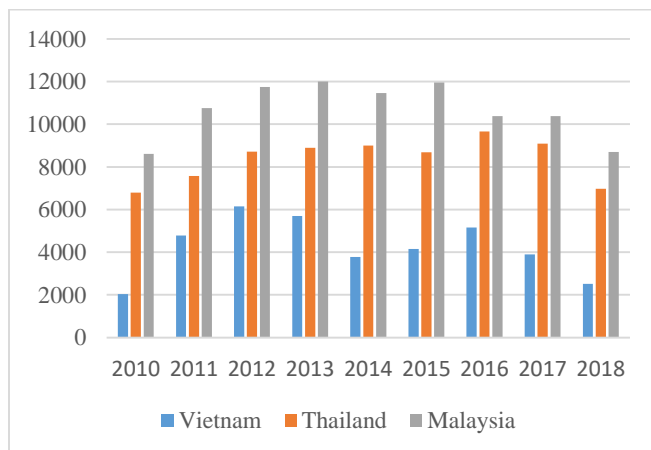


Figure 1. Comparing enterprises applying ISO 9001 in Vietnam, Thailand and Malaysia

Meanwhile, productivity and quality are important to the economy, productivity and quality improvement are a strategic goal in Vietnam's economic development plans and programs. On May 21, 2010, the Prime Minister issued Decision No. 712 / QD-TTg approving the National Project "Improving productivity and quality of products and goods of Vietnamese enterprises by 2020". However, the number of Vietnamese enterprises applying TQM through ISO 9001 is still low.

Every year, Vietnam has the policy to encourage enterprises applying quality management system as well as enhancing productivity and product quality. However, the quantity of studies in total quality management was limited. Most of them are general judgments without quantitative analysis with statistics to have a more specific view. Moreover, service sector is growing strongly, contributing greatly to the economy. While manufacturing companies bring more environmental pollution, promoting the development of service companies is a sustainable solution for developing countries. Hence, it is necessary to enhance TQM application in this field. Therefore, the purpose of this paper is to analyze total quality management applications and the impact of total quality management on business results.

## II. LITURATURE REVIEW

### A. Total Quality Management

Total quality management represents a movement which revolutionizes the way business is done in the industrialized world [2]. There are many points of the quality gurus about TQM definition.

According to [2] TQM is "an ongoing process whereby top management takes whatever steps necessary to enable everyone in the organization in the course of performing all duties to establish and achieve standards which meet or exceed the need and expectations of their customers, both external and internal".

In addition, TQM encompasses every aspect of the business or organization, not just the systems used to design, produce, and deploy its products and services. This includes all support systems such as human resources, finance, and marketing. TQM involves every function and level of the organization, from top to bottom [3]. Total quality management also means that management is responsible for developing the organization's vision (what it hopes to be at a point in the future), establishing guiding principles (a code of conduct for the organization and all of its employees), and setting the strategy and tactics for achieving the vision within the constraints of the guiding principles. In a TQM organization, the vision is pursued with input from an

empowered workforce that cooperates and collaborates with management. Therefore, TQM application have a positive effect to performance [3] [4] [5].

Based on the foregoing discussion, the factors of TQM were defined as the table below:

TABLE I. THE FACTORS OF TQM

Variables	Content
Context of the organization	the company understands the needs and expectations of its stakeholders in order to devise strategies appropriate to the organization's context [5][6]
Leadership	Top management is interested, responsible and very determined in implementing the quality management system [4][5] [6] [7]
Planning	Our company has a quality policy and a plan to achieve it[5][8][6]
Support	The members of the company all know about the TQM [4][5] [6] [7]. In our company, the information is widely and promptly disclosed to all employees [4][5] [6] [7].
Operation	All work in our company have standardized process [4][5] [6] [7]. All members of the company know and follow the standard procedures for their work [4][5] [6]
Performance evaluation.	All members of the company know and follow the standard procedures for their work [4][5] [6] Our company evaluates internal periodically [4][5] [6] [7]
Improvement	Our company regularly finds nonconformities to take preventive and corrective action for arising problems [4] [6] [7]. Our company has continuous improvement activities [4] [6] [7]

### B. Nine key service industries

According to the City General Statistics Office, the codes for the nine key industries are shown in Table 2 as follows:

TABLE II. NINE KEY INDUSTRIES

Industries code	Industries name
G	Commercial and repair of motor vehicles
H	Transportation and storage
I	Accommodation and food services
J	Information and communication
K	Finance, banking and insurance activities
L	Real estate business
M	Professional activities, science and technology
P	Education and training
Q	Medical

Source: General Statistics Office in Ho Chi Minh City, 2019

### C. The relationship between Total Quality Management and firm performance.

Total quality management (TQM) is one of the popular management methods in many developed countries, so there is quite a lot of research related to this field.

A research in US used the data from a cross-sectional mail study conducted to investigate TQM, just-in-time purchasing (JITP) and the performance of firms operating in the 48 contiguous states of the US that have implemented TQM and

JITP techniques. A large sample size was required to obtain reliable and valid research results. The result showed that quality management had a positive impact on financial and marketing performance [9].

Another study aimed to explore the relationship between quality management (QM) practices, quality performance and financial performance of the manufacturing firm. The authors identified and classified critical QM practices into categories. The empirical data was collected from a questionnaire-based survey of 152 Indian manufacturing companies. After the data was obtained, the scales were purified using loading values and composite reliability. The result showed a positive relationship between QM practices, quality performance and financial performance [10].

Next, a research project was carried out in 72 Spanish service companies to focus on TQM implementation. The authors formulated two measurement models. The first model includes the TQM practices, while the second contained the performance outcomes. The dimensions of the TQM factors were the quality practices of top management, process management, employee quality management, customer focus and employee knowledge and education. The performance outcomes represented by financial performance, customer satisfaction, product/service quality performance and operational performance [11].

In Vietnam, there were some studies that tested the impact of TQM on firm performance.

A study developed conceptual framework about total quality management of Hanoi construction companies. This framework was used not only to evaluate the practices of TQM, but also examine the relationship between TQM and organizational performance. The sample of the study includes construction companies in Hanoi. The authors concluded organizations which implemented the activities of TQM would positively increase their performance[12].

This study seeks to investigate the relationship between quality management practices and sustainability performance. Authors collected data from enterprises in Vietnam from July 2016 to March 2017 and there were 144 valid responses. The results indicated that quality management practices had mixed impacts on economic performance and environmental performance, while showing positive impact on social performance. Moreover, the study found significant moderating effects of three contextual factors on the relationship between quality management practices and sustainability performance[13].

Resently, the research identified the relationship between TQM practices and the performance in Vietnamese enterprises. The author surveyed 211 Vietnamese enterprises and the estimation results proposed that non – financial performance played a vital role as a full mediator in the relationship between TQM practices and financial performance in the Vietnam context [14].

From the above studies, it shows that almost studies in Vietnam using small data. In this study, with a large number of surveyed enterprises, authors hoped to make a great contribution with significant results.

## III. MODEL AND METHDOLOGY

Based on survey data of 1525 businesses operating in 9 key service industries in Ho Chi Minh City, the paper

performed quantitative analysis with OLS regression analysis and descriptive statistics. The effectiveness of the factors of TQM applications were asked, assessing from 0 to 10 (0: No effective, 10: Very effective)

Analytical methods that aim to identify specific quality management activities and the impact of quality management on business results.

#### Descriptive statistics:

- ✓ Percentage of enterprises applying TQM in 9 service sectors
- ✓ Compare the average score of the TQM factors in 9 service sectors
- ✓ Compare the average revenue and profit of the companies which have TQM application and no apply.

#### OLS regression analysis:

Two dependent variables that represent the firm performance are REVENUE and PROFIT.

The research model can also be presented as mathematical equation as follows:  $Y_i = \beta_1 + \beta_2 X_i + u_i$

$Y_i$  is the dependent variable,  $X_i$  is the independent variables,  $u_i$  is the error, and  $\beta_1$  and  $\beta_2$  are the regression coefficients.

## IV. RESULTS

### A. Percentage of enterprises applying TQM in 9 service sectors

The result illustrates an average of 14.62% of businesses applying TQM in 9 service industries. In which, the lowest rate is 4.29% of the Education and Training sector; the highest proportion is for the health sector (32.61%) and next number is the motor vehicle repair and trading sector (21.72%).

TABLE III. PERCENTAGE OF ENTERPRISES APPLYING TQM

Industries	Total	TQM application	
		Quantity	Percent (%)
Commercial and repair of motor vehicles	442	96	21.72
Transportation and storage	162	23	14.20
Accommodation and food services	188	20	10.64
Information and communication	158	11	6.96
Finance, banking and insurance activities	88	5	5.68
Real estate business	126	8	6.35
Professional activities, science and technology	245	35	14.29
Education and training	70	3	4.29
Medical	46	15	32.61
<b>Total</b>	<b>1.525</b>	<b>223</b>	<b>14.62</b>

Source: Authors' summary

### B. Analyzing the effectiveness of the application of TQM in 9 service sectors

In order to see the effectiveness of the TQM in the companies, the authors calculated the average score of the factors surveyed enterprises applying TQM (on a scale of 10).

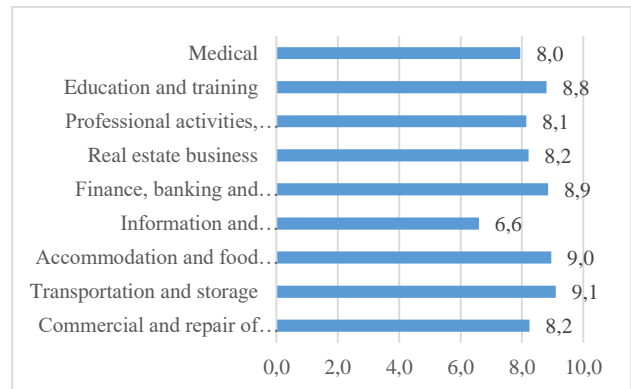


Figure 2. The average score of the TQM factors

Source: Authors' summary

From the analysis results show that in the enterprises applying TQM in 9 service sectors, the enterprises in transport and storage industry have the highest application efficiency with an average score of 8.98 (on a scale of 10 points); while professional, scientific and technological activities were the lowest with an average score of 7.8.

### C. Compare the average score of each factor of TQM in nine service industries

The authors calculated the average score of each factor of TQM to define weak factors which need to be improve in nine service industries. The results show in Table 4 and Table 5.

TABLE IV. THE AVERAGE POINT OF EACH FACTOR

	Context of the organization	Leadership	Planning
Commercial and repair of motor vehicles	8.0	8.5	8.1
Transportation and storage	9.0	9.5	9.1
Accommodation and food services	8.9	9.2	8.7
Information and communication	6.5	6.6	6.5
Finance, banking and insurance activities	8.8	8.5	9.3
Real estate business	8.3	8.5	8.2
Professional activities, science and technology	8.3	8.6	7.9
Education and training	8.7	9.0	8.7
Medical	7.8	8.5	8.0

TABLE V. THE AVERAGE POINT OF EACH FACTOR (CONTINUE)

	Support	Operation	Performance evaluation	Improvement
Commercial and repair of motor vehicles	8.1	8.3	8.3	8.4
Transportation and storage	9.0	9.1	9.1	9.2
Accommodation and food services	8.8	9.1	8.9	9.1
Information and communication	6.5	6.7	6.6	6.7
Finance, banking and insurance activities	8.6	9.0	8.6	9.3

Real estate business	7.8	9.2	8.1	8.1
Professional activities, science and technology	8.2	8.2	8.0	8.0
Education and training	8.7	9.0	8.7	9.0
Medical	8.1	8.3	7.8	7.5

Source: Authors' summary

The author analyzed the specific average score of the seven factors surveyed, the results are shown in Table 4 and Table 5. The results show that the highest score is the "Leadership" in transport and storage industry (9.5); the lowest score is the "support" of information and communication (6.5).

#### D. Regression results.

To better understand the effect of the application of TQM on business performance, the author uses the regression method with the dependent variable profit and revenue. The following is the detailed results of the regression analysis. The results of the analysis with the dependent variable REVENUE are given in Table 6.

TABLE VI. REGRESSION REVENUE AS DEPENDENT VARIABLE

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.6185	0.0223	206.91	0.000	
TQM	0.06518	0.00481	13.54	0.000	1.00

Source: Authors' summary

Specifically, the average revenue of all businesses with TQM activities is higher than the average revenue of enterprises that do not have TQM activities. In state-owned enterprises that use TQM, the average turnover is 1.62 times that of the state-owned enterprises that do not apply TQM, while this rate in foreign-invested enterprises is 1.42 times, and in private enterprises it is 1.01 times.

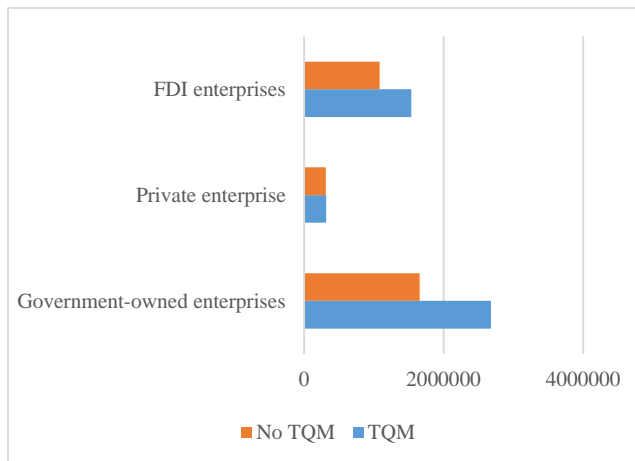


Figure 3. The relationship between revenue and TQM application

Source: Authors' summary

Next, with the dependent variable being profit, the regression results also show that TQM activities have a great influence on the profitability of enterprises.

TABLE VII. REGRESSION PROFIT AS DEPENDENT VARIABLE

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	67371	16743	4.02	0.000	
TQM	13307	3622	3.67	0.000	1.00

Source: Authors' summary

However, this influence is different from revenue in different types of businesses. If the revenue does not have much difference in private enterprises, the profits in these enterprises are very different, specifically the profit in non-state enterprises that apply TQM activities is 10.58 times higher than the profit at non-state enterprises does not apply TQM activities.

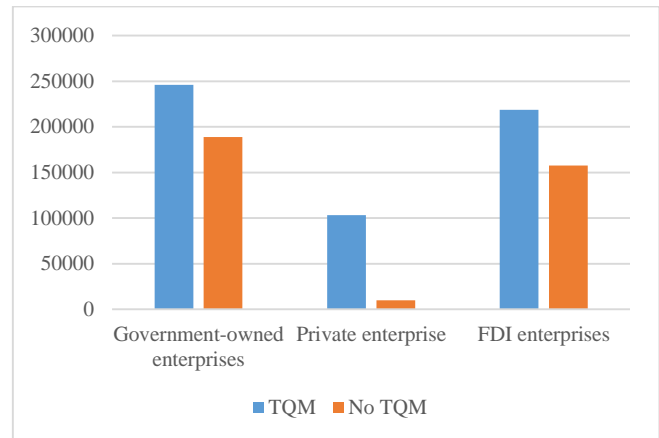


Figure 4. The relationship between profit and TQM application

Source: Authors' summary

## V. CONCLUSION AND SOLUTIONS

This study applied OLS regression to determine the impact of TQM on firm performance and analyze TQM application in 9 service industries. The results show that there was an average of 14.62% of businesses which have applied TQM. In which, the lowest rate is 4.29% of the Education and Training sector; the highest proportion was found in the Health sector (32.61%).

In general, "Support" is the one with the lowest score in all 9 service industries. According to experts, the improvement activities in enterprises are still coping, and there is no activity to check the results of the improvement and compare the effectiveness with the improvement.

Besides, the Standardized Processes factor is also the weak factor in 9 services sectors. This is a problem of businesses that needs attention. Standardization of processes is very important in the service. The characteristics of the service are heterogeneous, the quality of service provided by different employees at different times is not the same. Therefore, in order to ensure uniform service quality and customer satisfaction, standardization of the process is the solution that experts recommend service businesses to use. However, this factor is the lower among 10 factors surveyed, so service enterprises need to overcome this problem.

*Limitation of the topic:* Although the study has achieved some results, the research still has some limitations. The study only analyzes data in a specific period of time, has not yet analyzed data to see the fluctuation of the dependent variable over time. The thesis has just verified the results by

logit regression model but has not compared the results with other models. The study was conducted only in nine service industries in Ho Chi Minh City but not in other localities in Vietnam. The authors hope to carry out further studies to improve the above limitations.

#### REFERENCES

- [1] International Organization for Standardization, "The ISO survey," 2018. [Online]. Available: <https://www.iso.org/the-iso-survey.html>.
- [2] W. J. Miller, "A working definition for total quality management (TQM) researchers," *J. Qual. Manag.*, vol. 1, no. 2, pp. 149–159, 1996.
- [3] D. L. Goetsch and S. Davis, *Quality management for organizational excellence: introduction to total quality*, Seventh Ed. Pearson Education Limited, 2013.
- [4] J. Carlos Pinho, "TQM and performance in small medium enterprises," *Int. J. Qual. Reliab. Manag.*, vol. 25, no. 3, pp. 256–275, Mar. 2008.
- [5] D. I. Prajogo and A. S. Sohal, "The integration of TQM and technology/R&D management in determining quality and innovation performance," *Omega*, vol. 34, no. 3, pp. 296–312, Jun. 2006.
- [6] ISO, "ISO 9001:2015 Quality management systems — Requirements," 2015.
- [7] S. M. Yusof and E. Aspinwall, "TQM implementation issues: review and case study," *Int. J. Oper. Prod. Manag.*, vol. 20, no. 6, pp. 634–655, Jun. 2000.
- [8] M. Feng, M. Terziovski, and D. Samson, "Relationship of ISO 9001:2000 quality system certification with operational and business performance," *J. Manuf. Technol. Manag.*, vol. 19, no. 1, pp. 22–37, Dec. 2007.
- [9] H. Kaynak, "The relationship between total quality management practices and their effects on firm performance," *J. Oper. Manag.*, vol. 21, no. 4, pp. 405–435, 2003.
- [10] S. Parvadavardini, N. Vivek, and S. R. Devadasan, "Impact of quality management practices on quality performance and financial performance: evidence from Indian manufacturing companies," *Total Qual. Manag. Bus. Excell.*, vol. 27, no. 5–6, pp. 507–530, 2016.
- [11] C. Jaca and E. Psomas, "Total quality management practices and performance outcomes in Spanish service companies," *Total Qual. Manag. Bus. Excell.*, vol. 26, no. 9–10, pp. 958–970, 2015.
- [12] A. D. Nguyen, C. H. Pham, and L. Pham, "Total Quality Management and Financial Performance of Construction Companies in Ha Noi," *Int. J. Financ. Res.*, vol. 7, no. 3, May 2016.
- [13] M. Nguyen, A. Phan, and Y. Matsui, "Contribution of Quality Management Practices to Sustainability Performance of Vietnamese Firms," *Sustainability*, vol. 10, no. 2, p. 375, Jan. 2018.
- [14] T. M. D. Pham, "On the relationship between total quality management practices and firm performance in Vietnam: The mediating role of non-financial performance," *Manag. Sci. Lett.*, pp. 1743–1754, 2020.



# Synthesis of Zinc Oxide Nanoparticles and Their Antibacterial Activity

Thi Duy Hanh Le

Department of Chemical Technology  
Faculty of Chemical and Food Technology  
HCMC University of Education and Technology  
Ho Chi Minh city, Viet Nam  
duyhanhle@hcmute.edu.vn

Khanh Son Trinh

Department of Food Technology  
Faculty of Chemical and Food Technology  
HCMC University of Education and Technology  
Ho Chi Minh city, Viet Nam  
sontk@hcmute.edu.vn

**Abstract**— Zinc oxide nanoparticles (ZnO NPs) have received significant interest in bioengineering. Herein, ZnO NPs synthesis was performed via sol-gel method using precursors including zinc acetate salt and sodium hydroxide, followed by calcination of precipitated part at 180°C. The synthesized ZnO NPs properties in terms of phase composition, size and shape were then characterized by X-ray diffraction (XRD), ultraviolet–visible spectroscopy (UV-vis) and transmission electron microscopy (TEM). To test antibacterial ability of the synthesized ZnO NPs, glass pieces were covered with suspension containing different concentration of ZnO NPs synthesized, afterwards put into *Escherichia coli* (E.coli) culture after 24 hours of incubation at 37°C. The synthesized ZnO NPs characterized by TEM were mainly rod shape with particles size ranging 30-100 nm; moreover, UV-vis spectrum of the ZnO NPs suspension presented an absorbance peak at 371nm. Furthermore, antibacterial activity of them assessed throughout visualization of *E. coli* live was found to be influenced on dosed used.

**Keywords**—Zinc oxide nanoparticles, sol-gel methods, coated layer, antibacterial property.

## I. INTRODUCTION

Bacterial adhesion and colonization to the surface of materials or host have linked to many undesired problems in food processing and storage, water treatment, medical devices that have been well reported [1], [2]. To solve this problems, many materials and techniques that have been proposed and used for anti-bacterial comprised of inorganic nanomaterials, organic compounds as well as coating layer [3], [4]. Among inorganic nanomaterial classification, metal oxide nanoparticles such as titanium oxide, silver oxide, copper oxide, magnesium oxide, zinc oxide have been received a great interesting of research as well as emerging industrial use owing to stability in different environment, ease of fabrication at low temperature; moreover, they can reduce side effects of antibiotic use [5].

Zinc oxide showed absence of toxicity for human at the concentration of 5g/kg body weight. They are therefore used for cosmetics and drugs due to their peculiar chemical and physical properties [6]. To develop their applications, antibacterial activity of zinc oxide nanoparticles has been recently investigated [7], although the process underlying their antibacterial effect do not completely understood yet. Some theories have been proposed that both disruption of cell membrane due to interaction of nano-size of zinc oxide (ZnO NPs) with bacteria surface and release of reactive oxygen species (ROS) inducing hydroxyl radical are thought to cause toxic effect on bacteria [8], [9].

Often, antibacterial ability of ZnO NPs significantly

depends on their physicochemical properties in terms of size and shape and dose [10], [11]. Meanwhile, antibacterial properties of ZnO NPs mentioned above have been controlled by synthesized methods and precursors used.

As previous research reported, ZnO NPs can be a potential candidate for antibacterial application so far [12]. The study aims to strengthen evidence of antibacterial activity of ZnO NPs synthesized from zinc acetate dehydrate precursor via the sol-gel method. In particular, properties of synthesized ZnO NPs are fully characterized in terms of size and shape by different techniques. Then, the effect of ZnO NPs on antibacterial is evaluated through modified disc diffusion examined on *Escherichia coli*. Furthermore, the present work will be considered as a primary step to develop further research of the ZnO NPs applications for food technology and biomedical use.

## II. MATERIALS AND METHODS

### A. Materials

Zinc acetate dihydrate ( $\text{Zn}(\text{CH}_3\text{COOH})_2 \cdot 2\text{H}_2\text{O}$ ) was purchased from Merck (Germany), while other chemical including sodium hydroxide (NaOH), hydrochloric acid (HCl) and carboxy methyl cellulose (CMC) were of reagent grade and ordered from Xilong (China).

*Escherichia coli* (E. coli – B482) was provided by the Vietnam Type Culture Collection (VTCC), Hanoi University of Science and Technology (Vietnam). Agar and nutrient broth were obtained from HiMedia Ltd. (India).

### B. Synthesis of zinc oxide nanoparticles

Synthesis of ZnO nanoparticles were performed via the sol-gel method. Briefly, 5.0 g zinc acetate dihydrate was dissolved in 25ml of deionized (DI) water at 50°C and stirred for 30mins to get a transparent solution, followed by slowly adding 3.65 ml of sodium hydroxide solution 12.5 M into the solution to occur a white precipitation of zinc hydroxide. Afterwards an obtained suspension was centrifuged at 8000 rpm for 10 minutes to collect precipitated solid. The collected precipitation was triply re-suspended in DI water, then centrifuged again to remove any trace of base. Consequently, the precipitation was dried at 80°C for 3 hours before calcined in furnace (Nabertherm 1400, Germany) with heating rate 7°C/min to 180°C and keep this temperature for 1 hour to obtain zinc oxide powder.

### C. Sample preparation and antibacterial test

3×3 cm of glass samples (Viglacera, Vietnam) were washed with deionized water and HCl 0.5 M to remove dusts

adhered on surface and increase adhesion of membrane containing zin oxide NPs in following stage. Then, the glass samples were immersed in DI and NaOH 10% to neutralize any trace of acid, followed by immersion in DI water for 15 minutes. Subsequently, the samples were dried at 100°C for 1 hour before covering a layer comprised of zinc oxide NPs.

ZnO NPs were dispersed in DI water at different concentration (particles concentration was at 0.5; 1.0; 2.5; and 5.0 mg/ml, respectively). Then, CMC was slowly added into the NPs suspension up to 0.1% wt. After, the suspension was vigorously stirred for 60 minutes at 60°C to achieve milky color.

Glasses after pretreatment was covered by 200 µl of suspension prepared above and dried at room temperature, followed by calcination at 600°C in furnace (Nabertherm 1400, Germany) with heating rate 3°C/min. Then, surface glass samples covered with NPs were sterilized at 121°C for 15 minutes before antibacterial testing.

Name codes and composition of glasses covered with NPs were given in table 1.

TABLE I. NAME CODES AND COMPOSITION OF COATING LAYERS CONTAINING ZNO NPs

Sample name	CMC, %	ZnO NPs, mg/ml
M <sub>1</sub>	0.1	0.5
M <sub>2</sub>	0.1	1.0
M <sub>3</sub>	0.1	2.5
M <sub>4</sub>	0.1	5.0

#### D. Methods

##### 1) Characterization of Zinc oxide NPs

Chemical and phase composition of powder after calcining were characterized by X-ray diffraction (Bruker-AXS: D8 ADVANCE, Germany) using Ni filtered Cu-K $\alpha$  generated at 15 KV). X-ray diffraction (XRD) pattern of synthesized powder was recorded in the diffraction angle ranging from 28° to 70°. Furthermore, the information of crystalline size could be calculated from peaks of the obtained XRD pattern following the Scherrer's equation as described:

$$D(2\theta) = \frac{K \times \lambda}{\beta \times \cos \theta} \quad (1)$$

Where  $\lambda$  is the wavelength of the X-ray beam applied ( $\lambda=0.15406\text{nm}$ );  $\beta$  is the line broadening measured from the width at half maximum intensity of peak; the constant K is as the function of shape factor that has a variety of value, however K can has a value of 0.94 [13];  $\theta$  is the Bragg angle.

Morphology and size of zinc oxide were examined by transmission electron microscopy (TEM) with a JEM 1400 (JOEL-Japan). The synthesized powder was dispersed in ethanol at 100 µl/ml before examining.

UV-vis absorption of zinc oxide suspension was recorded by an automated UV-vis spectrometer UH5300 HITACHI (Japan) in the range of wavelength from 250 to 700 nm to strengthen evidence of zinc oxide nanoparticles formation. Suspension containing ZnO at 100µg/ml in DI water was used to examine UV spectra.

Scanning electron microscope (Hitachi, Japan) associated with energy-dispersive X-ray analysis (EDS) were used to

detect the presence of zinc oxide particles on the surface of glass before testing antibacterial measurement.

##### 2) Antibacterial assay

Growth condition including 0.65 g of nutrient broth, 50ml of DI water and 1.10 g of agar was prepared and then sterilized at 121°C for 15 minutes as well.

Antibacterial property of ZnO NPs was performed using a modified disc diffusion method. Briefly, the glass covered with NPs samples were placed into petri plates. 20µl of E.coli suspension (10<sup>4</sup> CFU/mL) were uniformly spread into the glass surface and dried under hood at 37°C for 8 hours to vaporize free water. Afterwards 1.0 ml of the growth condition was poured on the top of sample surface adhered E. coli. Finally, the petri plates were cultured under hood at 37°C for 24 hours before testing.

Optical microscope was used to visualize viable E. Coli before and after culture time on the glasses with and without a layer containing ZnO NPs. Each glass sample of experiment was performed in triplicate.

### III. RESULTS AND DICUSSION

#### A. Characterization of ZnO NPs

The XRD pattern (Figure 1) demonstrates presence of clear diffraction peaks at 31.82°; 34.48°; 36.31°; 47.60°; 56.62°; 62.92°; 67.95° and 69.12°, respectively. These peaks correspond to the wurtzite structure of zinc oxide (JCPDS 01 – 089 – 1397 index) in which zinc atoms located in the tetrahedral sites of hexagonal close packed shape of oxygen atoms [14].

As can be seen from the Figure 1, no peak attributes to possibility of impurities that can be existed in the obtained powder. In other words, the result shows that the synthesized ZnO is to be high purity. Moreover, the sharp peaks of pattern imply that zinc oxide synthesized is high crystallized and oriented, as previously reported [13], [15].

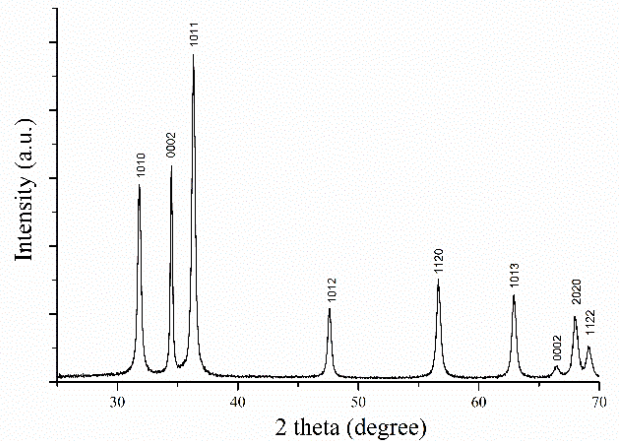


Figure 1. XRD pattern of the powder synthesized by sol-gel method

The UV-visible absorption spectrum (Figure 2) of ZnO suspension describes an absorption peak at 371nm that belongs a characteristic band of ZnO; moreover, the UV-vis result does not show any peak in the examined range. Yet, this result strongly supports the XRD analysis in the purity of ZnO synthesized.

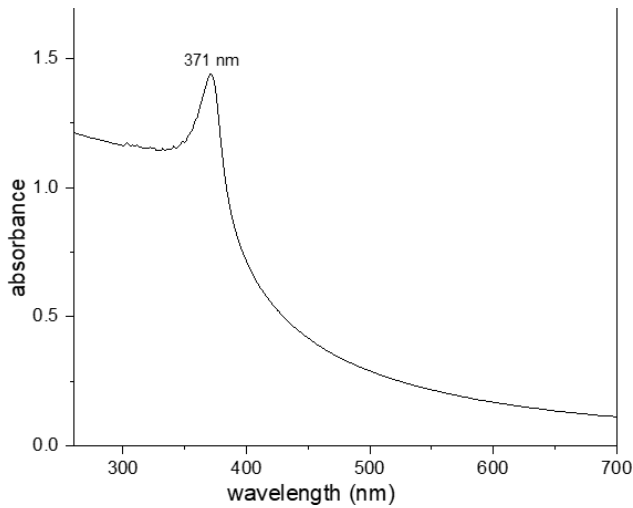


Figure 2. UV-vis absorption spectrum of synthesized the suspension at 100 $\mu$ g/ml of ZnO NPs.

UV-vis absorption band of ZnO nanoparticle depends on their size; however, its value is often in range of 330-380nm [14], [15]. Here, the appearance of peak at 371nm can provide an evidence of ZnO NPs.

TEM micrograph Figure 3 shows that synthesized ZnO particles mostly exhibit both round and rod shapes and have broad distribution. ZnO NPs particle size, which is based on their length and diameter, is estimated to range from 30 to 150nm.

TEM results are in good agreement with UV-vis and XRD results, testifying to the success of ZnO NPs formation by the sol-gel process that have been reported in many previous works.

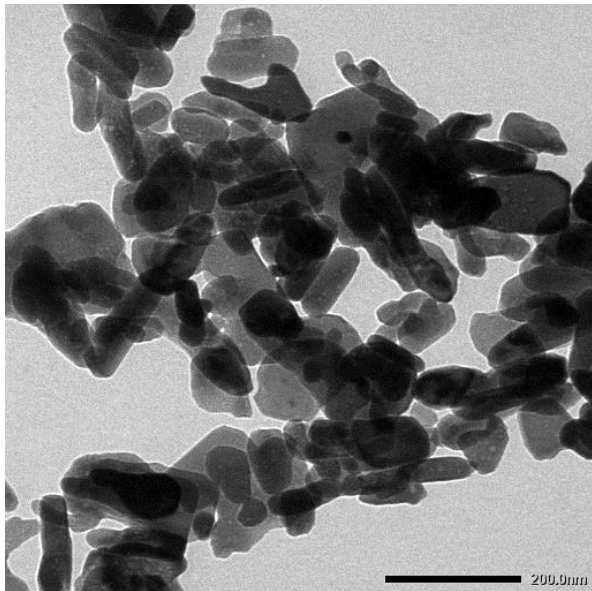


Figure 3. Morphology and size of ZnO-NPs studied by transmission electron microscopy

The Table 2 demonstrates different crystalline domain size of ZnO NPs calculated according to the Debye –Scherrer equation. The domain size is based on three high intensity peaks at 31.82, 34.84 and 36.31° of the XRD pattern. As calculated, the average crystalline domain size is about 27.3 nm.

TABLE II. CRYSTALLINE DOMAIN SIZE OF ZnO NPs CALCULATED TO THE DEBYE – SCHERRER EQUATION

$2\theta$ (°)	$\beta$ (radian)	D (nm)	$\bar{D}$ , nm
31.82	0.3745	23.5	27.3
34.84	0.2403	35.7	
36.31	0.3796	22.6	

There is a quite difference between TEM and XRD results in ZnO NPs particle size. Indeed, the calculation of size based on the XRD peak indicates crystalline domain size; therefore, the crystalline domain size may not be the same with particle size and a grain. It means that particles can contain multiple crystalline domains [15]. As a consequence, it is reasonable that particles size measured by TEM observation is larger than this size calculated based on diffraction peaks.

#### B. Characterization of ZnO NPs covered on the surface of glasses

The SEM micrographs ((Figure 4 A, B ) show complete difference in surface morphology of glass sample and glass sample covered with layer consisting of ZnO NPs, respectively. Particularly, a quite smooth surface (figure 4A) is observed in the surface of glass without a covered layer whilst the sample covered with suspension comprising of ZnO NPs displays a rough surface owing to a presence of ZnO NPs adhered to the glass surface after heat treatment. However, the layer containing ZnO NPs -covered glass surface shows an un-oriented and inhomogeneous of many large areas caused by nanoparticles aggregation during calcination.

The formation of nanoparticle aggregation on the surface of glass can be explained because of shrinkage of CMC layer containing ZnO NPs that results the shrinkage and then decomposition of CMC under the calcination at a low heating rate. Furthermore, the evenly distributed coating thickness of ZnO NPs layer could be a factor that attribute to nanoparticles aggregation phenomenon.

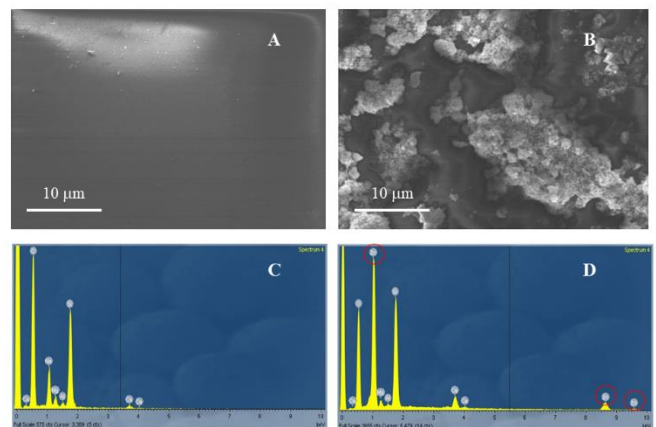


Figure 4. SEM micrographs of glasses without (A) and with (B) ZnO NPs layer on the surface, respectively; EDS spectra of glass sample (C) and glass sample covered with ZnO NPs (D). The red circles point to Zn peaks in the glass with ZnO NPs layer on its surface.

To determine the presence of ZnO NPs, elemental composition of both glass sample and glass covered with nanoparticle were assessed with EDS analysis; results are presented in figure 4 C, D. As can be seen in EDS measurement, the EDS spectrum of glass covered with ZnO



NPs shows clear peaks at the energy levels labelled as the presence of Zn (figure 4 D), whereas there is no signal of Zn elemental existence found in those spectrum of the pure glass sample (figure 4 C). The EDS results attribute to the successful deposition of ZnO NPs on surface of glass covered with nanoparticles.

### C. Antibacterial activity

The antibacterial activity of different concentration of ZnO NPs coated glasses is shown in figure 5 in which pure glass and glass coated with ZnO NPs were used as a negative control. White dots on glass surfaces shows a presence of *E. coli*.

The figure 5 shows an existence of *E. coli* after 24 hours of incubation in the surface of all glasses covered with ZnO NPs at any experimental concentration.

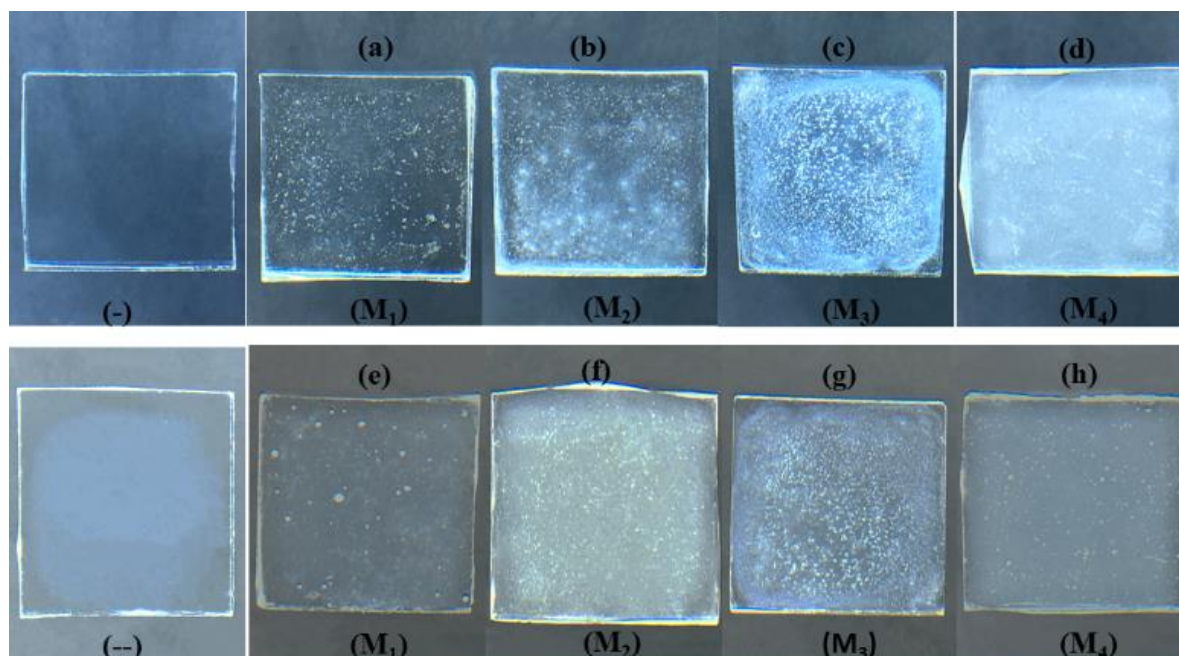


Figure 5: Measurement of anti *E. coli* of glasses covered with different ZnO NPs concentration at 0.5 (M<sub>1</sub>); 1.0 (M<sub>2</sub>); 2.5 (M<sub>3</sub>); 5.0 (M<sub>4</sub>) mg/ml, respectively. Images (a- d) were immediately taken after spreading *E. coli*; And images (e-h) were assessed after 24hours of incubation at 37° C; (-): glass without ZnO NPs and (--) glass covered with ZnO NPs at 2.5mg/ml.

In general, the number of *E. coli* after 24 hours of incubation decreased in all sample in comparison with the beginning. In addition, antibacterial ability of coated glasses seemed to depend on concentration of ZnO NPs used. However, dose-dependent regulation did not show. In particular, when the concentration of ZnO NPs in a covered layer was at 0.5 mg/ml and 5.0 mg/ml, the reduction in number of *E. coli* after 24 hours was significantly observed, compared to the beginning. In contrast, a slight decrease in bacterial number was found in samples with used concentration at 1.0 and 2.5 mg/ml.

As reported in previous studies, two key factors including high surface area of nanoparticles and generation of active oxygen species when ZnO NPs disperse in environment could have been used to explain the mechanism of ZnO NPs antibacterial [12], [16]. Here, the aggregation of ZnO NPs on the surface after calcination reduced surface area of ZnO NPs, followed by decrease of interaction between ZnO NPs and bacterial as well as the dissolution of ZnO in environment. Moreover, decrease of glass transparency after coating ZnO NPs process might affect visualization of antibacterial test.

To clarify the effect of nanoparticle aggregation on antibacterial ability of ZnO NPs, combination of spin-coater and fast-heating should be used to improve the thickness and distribution of coating layer on the surface as well as boost the combustion of CMC containing in the coated layer in further studies.

## IV. CONCLUSION

It is concluded that zinc oxide nanoparticles have been successfully synthesized via the sol-gel technique again. In the conducted study, the modified-disc diffusion method was used to test antibacterial ability, which showed a primary result of toxic effect of the synthesized ZnO NPs on *Escherichia coli*. Herein, we can state that application of ZnO NPs in antibacterial relies on the amount of ZnO NPs used.

Further studies with other protocols of antibacterial test are necessary to strongly elucidate ZnO NPs antibacterial property as well as fully understand precise antibacterial mechanism of this material.

## REFERENCES

- [1] F. R. Mizan, I. K. Jahid, and S. Ha, "Microbial biofilms in seafood: a food-hygiene challenge," *Food Microbiol.*, vol. 49, pp. 41-55, 2015.
- [2] S. Galié, C. García-gutiérrez, E. M. Miguélez, C. J. Villar, F. Lombó, and G. Di Bonaventura, "Biofilms in the Food Industry: Health Aspects and Control Methods," *Front. Microbiol.*, vol. 9, no. May, pp. 1-18, 2018.
- [3] R. Dastjerdi and M. Montazer, "A review on the application of inorganic nano-structured materials in the modification of textiles: Focus on anti-microbial properties," *Colloids Surfaces B Biointerfaces*, vol. 79, no. 1, pp. 5-18, 2010.
- [4] A. Clayton *et al.*, "Impact of curcumin nanoformulation on its antimicrobial activity," *Trends Food Sci. Technol.*, vol. 72, pp. 74-82, 2018.
- [5] S. M. Dizaj, F. Lotfipour, M. Barzegar-Jalali, M. H. Zarrintan, and K.

- Adibkia, "Antimicrobial activity of the metals and metal oxide nanoparticles," *Mater. Sci. Eng. C*, vol. 44, pp. 278–284, 2014.
- [6] K. Yamaki and S. Yoshino, "Comparison of inhibitory activities of zinc oxide ultrafine and fine particulates on IgE-induced mast cell activation," *Biometals.*, vol. 6, pp. 1031–1040, 2009.
- [7] A. Akbar *et al.*, "Synthesis and antimicrobial activity of zinc oxide nanoparticles against foodborne pathogens *Salmonella typhimurium* and *Staphylococcus aureus*," *Biocatal. Agric. Biotechnol.*, vol. 17, pp. 36–42, 2019.
- [8] A. Sirelkhatim, S. Mahmud, and A. Seenii, "Review on Zinc Oxide Nanoparticles: Antibacterial Activity and Toxicity Mechanism," *Nano-Micro Lett.*, vol. 7, pp. 219–242, 2015.
- [9] Y. Xie, Y. He, P. L. Irwin, T. Jin, and X. Shi, "Antibacterial Activity and Mechanism of Action of Zinc Oxide Nanoparticles against *Campylobacter jejuni*," *Appl. Environ. Microbiol.*, vol. 77, no. 7, pp. 2325–2331, 2011.
- [10] B. Lallo, M. P. Abuçafy, E. B. Manaia, and L. A. Chiavacci, "Relationship Between Structure And Antimicrobial Activity Of Zinc Oxide Nanoparticles: An Overview," *Int J Nanomedicine.*, vol. 14, pp. 9395–9410, 2019.
- [11] S. Sharma, K. Kumar, N. Thakur, and S. Chauhan, "The effect of shape and size of ZnO nanoparticles on their antimicrobial and photocatalytic activities: a green approach," *Bull. Mater. Sci.*, vol. 3, pp. 20, 2020.
- [12] R. Dadi, R. Azouani, M. Traore, C. Mielcarek, and A. Kanaev, "Antibacterial activity of ZnO and CuO nanoparticles against gram positive and gram negative strains," *Mater. Sci. Eng. C*, vol. 104, no. March, p. 109968, 2019.
- [13] A. K. Zak, M. E. Abrishami, W. H. A. Majid, R. Yousefi, and S. M. Hosseini, "Effects of annealing temperature on some structural and optical properties of ZnO nanoparticles prepared by a modified sol – gel combustion method," *Ceram. Int.*, vol. 37, no. 1, pp. 393–398, 2011.
- [14] Z. L. Wang, "Zinc oxide nanostructures: growth, properties and applications," *Journal of Physics: Condensed Matter*, vol. 16, pp. 829–858, 2004.
- [15] S. Thomas and P. Bindu, "Estimation of lattice strain in ZnO nanoparticles: X-ray peak profile analysis," *J Theor Appl Phys*, pp. 123–134, 2014.
- [16] B. Zhang, L. Cui, and K. Zhang, "Dosage- and time-dependent antibacterial effect of zinc oxide nanoparticles determined by a highly uniform SERS negating undesired spectral variation," *Anal Bioanal Chem*, pp. 3853–3865 vol. 408, no. 14, 2016.

# A Study on Design and Fabrication of Concrete Pipe Cutting Machine

Cong Binh Phan  
*Faculty of Mechanical Engineering*  
*HCMC University of Technology and Education*  
 Ho Chi Minh City, Vietnam  
 binhpc@hcmute.edu.vn

**Abstract**— Based on the actual demand, the semi-automatic concrete pipe cutting machine (CPCM) has been designed to improve performance of the conventional system. Firstly, a conceptual design is presented to describe the working principle of the CPCM. Then, mathematical modelling is calculated to obtain the equipment parameter. Next, The CPCM is fabricated to test the performance. The experimental results indicate that the proposed CPCM can work smoothly and safety. Moreover, it can reduce labor cost and increase the product quality. Therefore, the proposed CPCM is met the requirement of market.

**Keywords**—concrete pipe cutting machine, CPCM, cutter design, experimental setup, PID controller

## I. INTRODUCTION

In concrete drainage pipe manufacturing technology, there are some standards dimension to reduce the number of mold types. However, the actual demand of pipe's length is very different from the standard size, it must be cut to satisfy the customer's requirement. Moreover, some defect products can be also reused by cutting the not good end. Some manufacturers employ the manual labor to perform this process. It is not effective and not safety for labor to cut concrete by the hand tools. Therefore, the study on cutting machine is necessary.

There are some reports which refer to pipe cutting technology. The structural analysis of planetary pipe cutting machine based ANSYS has been presented by Ju Yi *et al.* [1]. Pneumatically Operated Automatic Pipe-Cutting Machine is studied by N Shripad *et al.* in [2]. Ju Yi *et al* have shown the design of cutting head for efficient cutting machine of thin-walled stainless steel pipe in [3].

However, it is not convenient to apply the present works for cutting the concrete pipe due to heavy and large objects. It is difficult and dangerous to exploit in the realistic conditions. The CPCM is designed to solve these problem.

This paper reports the proposed semi-automatic concrete pipe cutting machine (CPCM) which has been designed to improve performance of the conventional system. Firstly, a conceptual design is presented to describe the working principle of the CPCM. Then, mathematical modelling is calculated to obtain the equipment parameter. Next, The CPCM is fabricated to test the performance. The experimental

results indicated that the proposed CPCM can work smoothly and safety. Moreover, it can reduce labor cost and increase the product quality. Therefore, the proposed CPCM is met the requirement of market.

## II. CONCEPTUAL DESIGN AND WORKING PRINCIPLE

### A. Structural configurations

The geometrical parameters of the concrete pipe are shown in Fig. 1. Here, L1, D3 and D4 are the length of concrete pipe, the inner and outer diameter, respectively. For dimension of flared end, L2 and L3 are the lengths and D1 and D2 are the outer and inner diameters of the pipe, respectively.

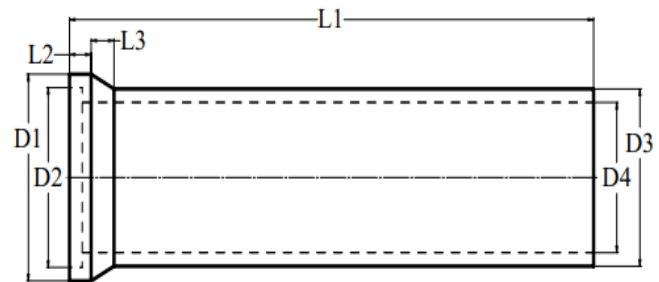


Fig. 1. Geometrical parameters of the concrete pipe

In this study, six sizes of concrete pipe from D300 to D1000 are measured to design the CPCM. The detail dimensions are shown in TABLE I.

TABLE I. DETAIL DIMENSIONS OF THE CONCRETE PIPES

Types of concrete pipe	Dimension (mm)						
	D1	D2	D3	D4	L1	L2	L3
D300	506	416	400	300	4060	80	137
D400	606	516	500	400	4060	80	137
D500	730	630	620	500	4080	118	159
D600	836	736	720	600	4080	109	168
D800	1080	976	960	800	4100	120	165
D1000	1376	1196	1180	1000	4120	170	180



The configuration of the given CPCM is shown in Fig. 2. There is an electric motor fixed to the input shaft of a gearbox. The remain gear fixed to the driving shaft rotates the driving wheels. The driven wheels are the follower of the pipe. Moreover, two supports are mated on the base frame to keep the concrete pipe stable. Finally, the cutting tool system is assembly on the base frame throughout the sliding bearing. A screw and nut mechanism driven by electric motor is employed to actuate the cutting tool along centerline of the concrete pipe

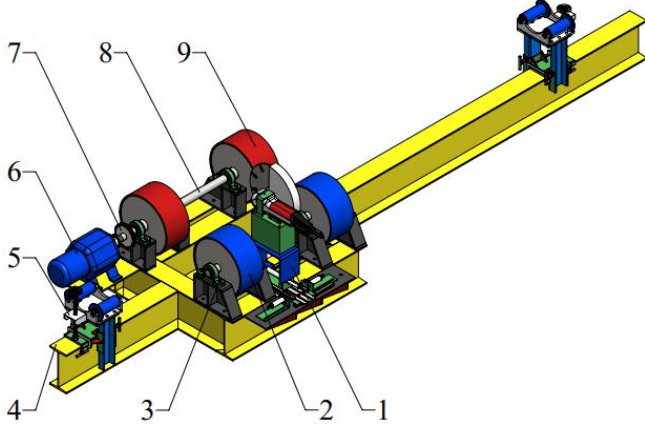


Fig. 2. Configuration of the proposed CPCM

1. Cutting tool system; 2. Drive flywheel; 3. Bearing; 4. Base; 5. Support; 6. Motor; 7. Gearing; 8. Driving shaft; 10. Driving wheel

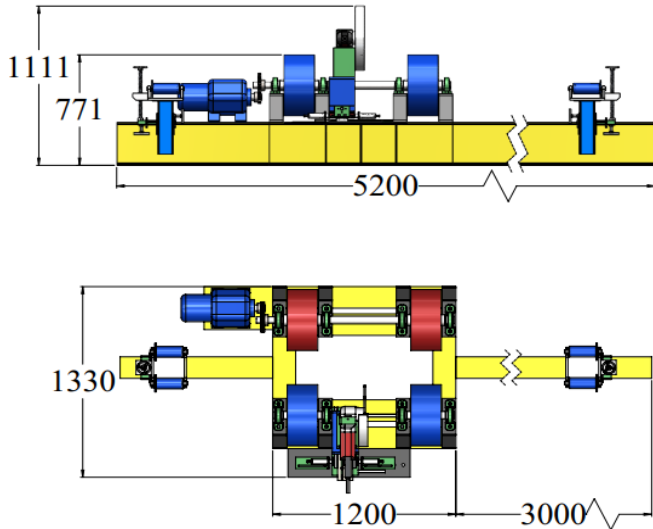


Fig. 3. Overall dimensions of the CPCM

Overall dimensions of the CPCM are shown in Fig.3. The length, the width and the height of the proposed CPCM are about 5200, 1330 and 1111 in mm, respectively.

### B. Working Principle

Firstly, a concrete pipe is placed on the four wheels. Then, the supports positions are adjusted to adapt with the suitable height. Next, the cutting tool is moved to cutting position. After setting ON the Test mode, the cutting line is checked to make sure the pipe stable. After checking the proper position, the Cut mode is set to cut through the inner of concrete pipe. Consequently, the concrete pipe is rotated and cut fully.

### III. MATHEMATICAL MODEL

The rotary motions of the concrete pipe are obtained by solving the differential equation which is applied Newton's second law in (1). The total torques acting on the pipe consist of the driving moment  $T_d$ ; friction torque  $T_f$  due to the wheel and cutting torque  $T_c$  due to the cutting tool.

$$J\ddot{\theta} = T_d - T_f - T_c \quad (1)$$

The cutting torque acting on the pipe are calculated in following equation:

$$T_c = F_c R_p \quad (2)$$

where  $R_p$  is the radius of the pipe;  $F_c$  is the cutting force which is generated by the cutting tool. Cutting force depends on two factors. These factors are the cutting speed  $v_c$  and the thickness of the pipe.

The friction torque of the system can be modeled and approximated with the following equation:

$$T_f = (T_c + (T_{br} - T_c)e^{(-c_v|\dot{\theta}|)})\text{sign}(\dot{\theta}) + f_v\dot{\theta} \quad (3)$$

where  $T_c$  is the Coulomb friction torque that opposes motion with a constant force at any angular velocity;  $T_{br}$  is the breakaway friction torque, which is the sum of the Coulomb and static frictions at zero velocity;  $f_v$  is the viscous friction coefficient and  $c_v$  is the transition approximation coefficient, which is used for the approximation of the transition between the static and the Coulomb frictions.

The drive moment is calculated in (4)

$$T_d = F_t R_p \quad (4)$$

where  $R_p$  is also the radius of the pipe;  $F_t$  is the tangent force that appears at the point of tangency between the pipe and the wheel. This factor can be calculated by the formula below:

$$F_t = \frac{T_{wheel}}{R_w} \quad (5)$$

where  $R_w$  is the radius of the wheel;  $T_w$  is the wheel torque. This component is depended on both the ratio and efficiency of the transmissions; it can be illustrated by the equation:

$$T_{wheel} = T_M i_g \eta \quad (6)$$

where  $T_M$  is the motor torque and can be obtained by doing experiment;  $i_g$  is the transmission ratio of the gearing and  $\eta$  is the overall efficiency.

Equation (7) is used for calculating the inertia of the pipe with annular section:

$$J = \frac{1}{2} m (R_{od}^2 + r_{id}^2) \quad (7)$$

where  $m$  is the mass of the pipe;  $R_{od}$  and  $r_{id}$  are the outer radius and the inner radius of the pipe, respectively.

The cutting velocity is obtained in (8)

$$v_c = \dot{\theta}_p R_p \quad (8)$$

Equation (9) is employed to calculate the rotational speed:

$$\dot{\theta}_{wheel} = \frac{v_c}{R_w} \quad (9)$$

where  $v_c$  is the cutting speed;  $R_w$  is the radius of the wheel.

The motor speed is obtained in following equation:

$$\dot{\theta}_M = \dot{\theta}_w i_g \quad (10)$$

The power of the motor is calculated by the equations:

$$P_M = \dot{\theta} T_M \quad (11)$$

#### IV. CONTROLLER

Schematic control diagram adjusting cutting tool position along the centerline of the concrete pipe is illustrated in Fig.4. Here, the setpoint is obtained by taking signal from position sensor measuring the displacement of concrete pipe along centerline. The output signal is obtained by another position sensor measuring the cutting tool positions. These signals are sent to a proportional-integral-derivative (PID) controller to obtain the control signal. Control system is built in Matlab/Simulink environment via data acquisition card PCI from Avantech Corp. The signals are obtained and sent to PCI card interfaced to Matlab/Simulink program. Based on the control signal, the motor is controlled adjust the cutter positions. Therefore, the cutting tool can track the concrete pipe positions.

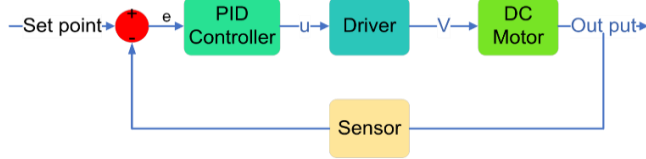


Fig. 4. Schematic control diagram for the cutting tool position

#### V. PROTOTYPE PERFORMANCE TESTS

##### A. Experimental Setup

Experimental setup for CPCM is presented in Fig. 5 and the drive mechanism is shown in Fig. 6. As shown in Fig. 5, the concrete pipe is placed on the four wheels. The two supports placed at the ends of the CPCM are adjusted manually to adapt with the suitable height. These supports keep the concrete pipe stable after finishing the cutting process. The pipe placing on the driving wheels is driven by the electric motor. A gearing illustrated in Fig. 6 is employed to reduce the driving speed and to increase the driving torque.

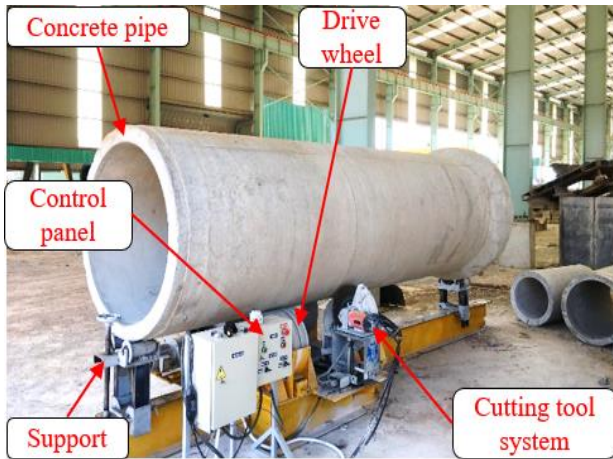


Fig. 5. Experimental setup for testing the CPCM

Some initial specifications of test rig are measured and shown in TABLE II and the setting parameters for doing experiment are presented in TABLE III.

As shown in TABLE II, a cutting tool employed in general cutting application is about 3000 rpm. For the driving system, a three-phase electric motor is selected which is about 1450 rpm in rotary speed. Based on the desired concrete pipe speed in Test mode, the transmission ratio is obtained and shown in Table. In this study, the concrete pipe D400 is employed to

setup experiment, and its specifications are also measured and indicated in Table.

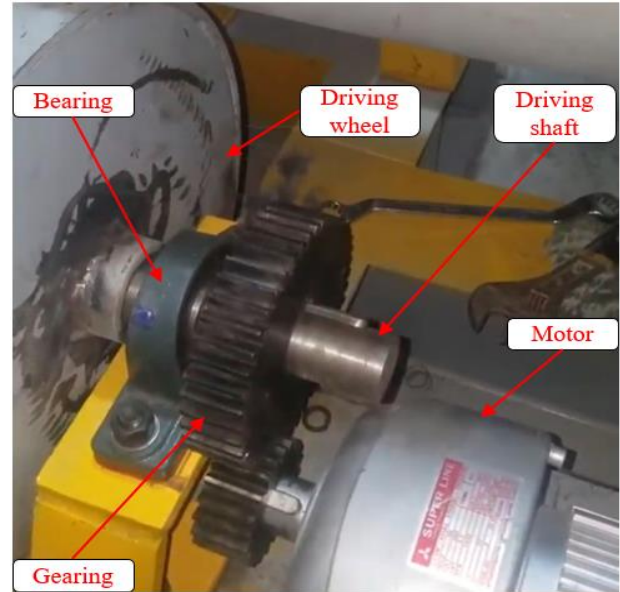


Fig. 6. Motor through a gearing connected to main shaft

TABLE II. SPECIFICATIONS OF TEST RIG

Specifications	Parameters
Cutting tool system speed [rpm]	3000
Driving motor speed [rpm]	1450
Transmission ratio $i_g$	1:180
Outside diameter of concrete pipe [mm]	600
Thicknesses of concrete pipe [mm]	100

The setting parameters for setting up experiment with D400 are obtained and presented in TABLE III. For the Test mode, the feed is selected by the realistic condition which is fast enough to save working time. The feed cutting mode is obtained by doing experiment to find out the optimum value for D400, which is satisfied both the safety factor and productivity. Based on the optimum feed cutting, the VFD is setup to control rotary speed of the concrete pipe. Here, the inverter frequency for the cutting mode is set about 1 Hz. Next, the gains in PID controller are obtained by trial and error method and shown in this Table. The given gains are not the optimum values; however, the respond of the system can be acceptable.

TABLE III. SETTING PARAMETERS FOR EXPERIMENTAL SET UP

Specifications	Parameters
Feed Test [m/min]	1
Feed cutting Mode [m/min]	0.5
Inverter frequency [Hz]	1
Proportional gain $K_p$	5
Integral gain $K_i$	1
Derivative gain $K_d$	0.01

##### B. Experimental Results

Fig. 7 shows the relative positions between the setpoint and the respond of the cutting tool. Because the concrete pipes

have form tolerance, the setpoints have been changing in time domain. PID controller is applied to adjust the cutting tool position. Then, the cutting tool can adapt to displacement of the concrete pipe. Therefore, the performance of CPCM is smooth and safe.

The experimental results in three cases of experiment are shown in TABLE IV. These results indicate that total maximum cutting time for D400 is about 12 minutes and the deviation error of cutting line is about 3 mm.

TABLE IV. RESULTS FROM THE EXPERIMENT





Specifications	Parameters		
	1st	2nd	3rd
Cutting time [min]	12	11	13
Deviation of cutting line [mm]	3	2.5	2

TABLE V presents performance comparison results between manual labor and CPCM. In the Hand Cutting Tool (HCT) method, the mean cutting time is measured directly at site where worker perform to cut the concrete pipe using hand tools only. It takes two peoples who work in not safe environment. Moreover, it takes so much time to move the concrete pipe using their hand, and cutting feed is not optimized by controller. Therefore, the total operation time is 5 times compared to that of CPCM. Here, the form tolerance of the cutting line in HCT is about 30 mm, whereas it is only maximum 3 mm in the CPCM performance. In addition, quality of cutting surfaces are visualized and compared to CPCM. As shown in Table VI, the surface roughness cut by CPCM is smoother than that of HCT.

TABLE V. PERFORMANCE COMPARISON FORM EXPERIMENTS

Specifications	HCT	CPCM
Mean cutting [min]	60	12
Deviation of cutting line [mm]	30	3
Surface roughness	OK	Very Good

TABLE VI. QUALITY VISUALIZATION COMPARISONS

Specifications	HCT	CPCM
Deviation of cutting line		
Surface roughness		

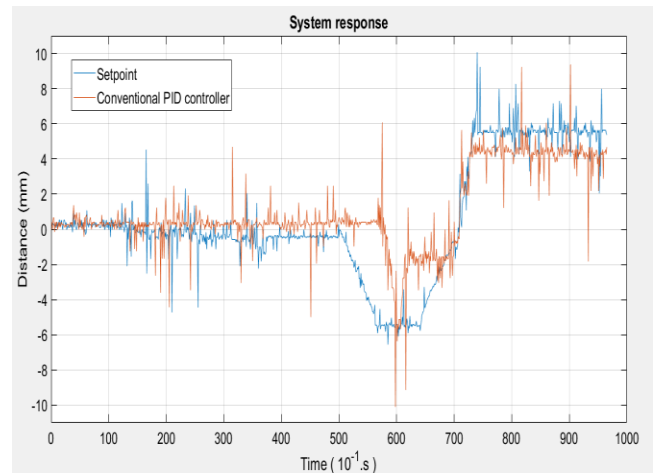


Fig. 7. Relative displacement of setpoint and cutting tool positions

## VI. CONCLUSIONS

This paper proposed a design of the CPCM which is applied to cut the heavy concrete piping into the desired dimensions. The PID control algorithm is employed to adjust the movement of the cutting tool automatically. The experimental results indicate that the proposed CPCM can work in the realistic conditions effectively. Some advantages can be found in the following:

- The productivity is about 500% increasing compared to which of manual labor
- The tolerance of cutting line is 90% reducing compared to manual labor
- The quality of cutting surface is improved clearly

Moreover, the labor cost is decreased significantly, and worker can operate CPCM in safe environment.

## REFERENCES

- [1] Ju Yi, Yingping Qian, Zhiqiang Shang, Zhihong Yan, Yang Jiao, Structure Analysis of Planetary Pipe Cutting Machine Based ANSYS, 2017.
- [2] Nimbalkar Shripad, Velanje Sagar, Patil Abhay, Varpe Pooja, Pneumatically Operated Automatic Pipe-Cutting Machine, Vol-2, Issue-2 2016.
- [3] Ju Yi, Yingping Qian, Zhiqiang Shang, Zhihong Yan, Yang Jiao, Design of Cutting Head for Efficient Cutting Machine of Thin-walled Stainless Steel Pipe, 2016.
- [4] A. A. Gurskiy, A. E. Goncharenko, S. M. Dubna, Algorithms for tuning of the coordinating automatic control systems, 2020.
- [5] Mu-Tian Yan, Pin-Hsum Huang, Accuracy improvement of wire-EDM by real-time wire tension control, 2004.
- [6] Mujtaba Jaffecy, Sohain Aslam, Moinuddin Ghauri, M. Shazad Khuram, Sikandar Rafiq, Subhan Khan, Real-time implementation of model predictive control on a 16-bit microcontroller for speed control of a dc motor, 2018.
- [7] Liuping Wang, Model Predictive Control System Design and Implementation using MATLAB, 2020.



# Optimal Day-ahead Energy Scheduling of Battery in Distribution Systems Considering Uncertainty

Ying-Yi Hong

Department of Electrical Engineering  
Chung Yuan Christian University  
Taoyuan City, Chung Li 32023, Taiwan  
yyhong@ee.cycu.edu.tw

Man-Yin Wu

Department of Electrical Engineering  
Chung Yuan Christian University  
Taoyuan City, Chung Li 32023, Taiwan  
mm1223s@yahoo.com.tw

Sheng-Huei Lee

Department of Electrical Engineering  
Chung Yuan Christian University  
Taoyuan City, Chung Li 32023, Taiwan  
garylee@uch.edu.tw

**Abstract**— One of the features in the modern smart grid is accommodation of different distributed generation resources and battery energy storage system (BESS). Because the renewable power generation is intermittent and uncertain, the BESS can help regulation of power generation, voltage and even system frequency. This paper explores a short term (24 h) energy scheduling of batteries in a power system considering uncertain photovoltaic (PV) power generations and loads. The cost of losses is minimized while both equality and inequality constraints are satisfied. The equality constraints consist of the power flow equations and the energy balance equations for batteries. The inequality constraints comprise the limits of power and energy of batteries. The states (charging and discharging) of batteries are set according to the tariffs of electricity. Due to the uncertainty of PV powers and loads, the problem is solved by two loops: the outer loop is implemented by point-estimation method while the inner loop deals with the 24 h optimal power flow considering deterministic PV generations and loads using the interior-point method. A 33-bus distribution system of 3.454 MW (peak load) was used to show the simulation results. One PV farm (1MW) and six BESS at different buses were considered. The deterministic and stochastic results obtained by the proposed method were discussed and compared.

**Keywords**-- Energy storage system, Point-estimation method, Scheduling, Uncertainty.

## I. INTRODUCTION

Modern smart distribution system can accommodate different distributed power generation resources (including renewable energy and micro-turbine generators, et al.) and the battery energy storage system (BESS) enabling bi-directional energy flow [1]. Electrical power generation from renewable energies is uncertain because their characteristic is intermittent [2]. The BESS can provide voltage support [3] and frequency regulation [4]. Specifically, the BESS can mitigate fluctuations of the output power from PV arrays and wind farms [5]. The BESS may also incorporate traditional thermal generators to reduce the oscillation of power outputs from renewable power sources [6, 7] and provide a seamless transition between islanding and grid-tied modes in a microgrid [8].

In order to incorporate the BESS with other conventional units, many works in the past have been proposed. Duggal and Venkatesh proposed a battery cycling and depth of discharge relation for utility-scaled BESS and presented a mathematical formulation for the short-term 24-h scheduling problem in conjunction with thermal generation [9]. Matthias et al. proposed a cooperative multi-area

optimization strategy to enable transmission system operators to dispatch/redispach interconnected networks securely, while reducing dispatch/redispach costs [10]. Schedules for storage devices, conventional- and renewable generation were obtained considering network constraints and ramping rates [10]. Luna et al. presented the modeling and design of a modular energy management system and its integration to a grid-connected battery-based microgrid. The scheduling model was a power generation-side strategy, which was solved by general mixed-integer linear programming [11]. Wu et al. presented a stochastic day-ahead scheduling of electric power systems with flexible resources including thermal units with up/down ramping capability, energy storage, and hourly demand response [12]. The Monte Carlo simulation was used for simulating random outages of generation units and transmission lines [12]. Degeilh and Gross formulated a scheduling optimization problem to determine the operational schedule of the controllable storage resources in coordination with the demands and the various supply resources, including the conventional and renewable resources, through the Monte Carlo simulation framework [13]. Jabr et al. presented a sparse formulation and solution for the multi-period OPF problem, which aimed at operating a storage portfolio via receding horizon control [14]. It computed the optimal base-point conventional generation and storage schedule taking the forecasted load and renewable generation, together with the constrained participation factors that dictate how conventional generation and storage will adjust to maintain feasible operation, into account [14].

According to the above existing works, it is essential to explore the short-term energy scheduling of BESS incorporating with conventional units, thermal generators. The above methods can be improved at least in two areas: (a) modeling of uncertainty: the works in [9-11] did not involve this issue caused by intermittent the renewable energy and load and (b) complexity: the works were explored by Monte-Carlo simulations, which take a long CPU time [12, 13], or complicated methods [14].

This paper investigates a short term (24 h) energy scheduling of BESS in a power system considering uncertain photovoltaic (PV) power generations and loads. The cost of losses is minimized while both equality constraints (the power flow equations and the energy balance equations for BESS) and inequality constraints (the limits of power and energy of BESS) are satisfied. The problem is solved by two loops: the outer loop is implemented by the point-estimation method (PEM) [15, 16] that considers random variables (PV and load) while the inner loop deals with the 24 h optimal power flow considering deterministic PV generations and loads using the interior-point method (IPM) [17].

The rest of the paper is organized as follows. Section II

This work is sponsored by Ministry of Science and Technology, Taiwan under the Grant MOST 108-2221-E-033-023 and MOST 109-3116-F-006-019 -CC1.

presents a detailed description and formulation of the problem. Section III presents the proposed method, which involves the PEM and IPM. Section IV presents the results of simulations using a 33-bus distribution system of 3.454 MW (peak load), 6 BESS and 1-MW PV farm. Section V draws conclusions.

## II. PROBLEM DESCRIPTION AND FORMULATION

### A. Problem Description

This paper explores a short term (24 h) energy scheduling of BESS in a distribution system considering uncertain photovoltaic (PV) power generations and loads. Some issues of this problem are described as follows:

(1) Uncertainty: Since the day-ahead study is conducted, uncertainty factor must be taken into account. The PV power and bus load are considered to be uncertain in this paper because they cannot be controlled by human being or engineer.

(2) Electricity tariffs: Different power markets have different electricity tariff mechanisms, which may be spot price, two/three tariffs in a day or uniform tariff. Tariff mechanisms have great impacts on charging/discharging schemes for the BESS that could regulate power flow to mitigate the power loss. This paper adopts two tariffs in a day.

(3) Operation constraints: Operation of the power system must be subject to the power flow equations, voltage and line flow constraints. Besides, the charging/discharging of BESS is constrained by its corresponding capacity (kW). To prolong the BESS life, the state-of-charge (SOC) of BESS must be within a proper range that is expressed by kWh herein.

(4) Charging/discharging strategy: The charging/discharging strategy of the BESS depends on many factors, such as tariffs and SOC. In order to minimize the loss cost, the BESS is charged (discharged) in the low (high) tariff period, corresponding to off-peak (peak) hours.

### B. Problem Formulation

Based on the above description, the studied problem can be formulated as follows.

$$f = \min CE_{loss} \quad (1)$$

where  $CE_{loss}(h)$  denotes the cost of energy losses at hour  $h$ . More specifically,

$$CE_{loss} = \sum_{\text{peak hour}} P_{loss}(h) \times C_{\text{peak}} + \sum_{\text{offpeak hour}} P_{loss}(h) \times C_{\text{off-peak}} \quad (2)$$

and  $C_{\text{peak}}$  and  $C_{\text{off-peak}}$  are the electricity tariffs of peak and off-peak loads, respectively. The variable  $P_{loss}(h)$  represents the real power loss at hour  $h$ .

The objective function in (1) and (2) must be subject to the inequality constraints related to all BESS as follows.

$$0 \leq E_{BESS}(n, h) \leq E_{BESS}^{rated}(n), \quad n \in N, h \in H \quad (3)$$

$$0 \leq P_{BESS}^{dis}(n, h) \leq P_{BESS}^{rated}(n), \quad n \in N, h \in H_d \quad (4)$$

$$0 \leq P_{BESS}^{ch}(n, h) \leq P_{BESS}^{rated}(n), \quad n \in N, h \in H_c \quad (5)$$

where  $E_{BESS}(n, h)$  and  $E_{BESS}^{rated}(n)$  signify the energy (kWh) of the  $n$ th BESS at hour  $h$  and its corresponding size (kWh), respectively.  $P_{BESS}^{ch}(n, h)$  and  $P_{BESS}^{dis}(n, h)$  denote the charging and discharging power (kW) of the  $n$ th BESS at hour  $h$ , respectively.  $P_{BESS}^{rated}(n)$  is the rating power of the  $n$ th BESS. Please note the charging and discharging periods will not occur at the same time.  $N$  is the set of BESS. The symbols  $H$ ,  $H_d$ ,  $H_c$  imply the sets of 24 h, hours for discharging and hours for charging, respectively. Let the set  $Nbus$  be the total buses. The power system is also subject to the following operational constraints:

$$V_i^{min} \leq V_i(h) \leq V_i^{max}, \quad i \in Nbus, h \in H \quad (6)$$

$$0 \leq |f_{ij}(h)| \leq f_{ij}^{max}, \quad i, j \in Nbus, h \in H \quad (7)$$

where  $V_i(h)$  is the voltage magnitude at bus  $i$  and hour  $h$ ;  $f_{ij}(h)$  is the line flow between buses  $i$  and  $j$  at hour  $h$ . The superscripts “max” and “min” are the maximum and minimum limits, respectively.

The objective function is also subject to equality constraints consisting of energy balance equations. Let the studied interval for two consecutive operating points be one hour. Then

$$E_{BESS}(n, h+1) = E_{BESS}(n, h) + \eta_c \cdot P_{BESS}^{ch}(n, h) \quad n \in N, (h+1) \in H_c \quad (8)$$

$$E_{BESS}(n, h+1) = E_{BESS}(n, h) - \eta_d \cdot P_{BESS}^{dis}(n, h) \quad n \in N, (h+1) \in H_d \quad (9)$$

where  $\eta_c$  ( $\eta_d$ ) is the efficiency for the BESS charging (discharging). The equality constraints comprise the power flow equations. For bus  $i$ ,

$$PG_i(h) - PD_i(h) - P_{BESS}^{ch}(i, h) = \sum_{j=1}^{Nbus} V_i(h) \cdot V_j(h) \cdot Y_{ij} \cdot \cos(\theta_{ij} + \delta_{ij}(h)) \quad (10)$$

$$PG_i(h) - PD_i(h) + P_{BESS}^{dis}(i, h) = \sum_{j=1}^{Nbus} V_i(h) \cdot V_j(h) \cdot Y_{ij} \cdot \cos(\theta_{ij} + \delta_{ij}(h)) \quad (11)$$

$$QG_i(h) - QD_i(h) = - \sum_{j=1}^{Nbus} V_i(h) \cdot V_j(h) \cdot Y_{ij} \cdot \sin(\theta_{ij} + \delta_{ij}(h)) \quad (12)$$

where  $PG_i(h)$ ,  $PD_i(h)$ ,  $QG_i(h)$ , and  $QD_i(h)$  are PV real power generation, real power demand, PV reactive power generation and reactive power demand at bus  $i$  and hour  $h$ , respectively.  $Y_{ij} \angle \theta_{ij}$  and  $\delta$  represent the element of bus admittance matrix and bus voltage phase angle, respectively.

## III. PROPOSED METHOD

In the studied problem,  $PG_i(h)$  and  $PD_i(h)$  are random variables.  $QG_i(h)$  can be assumed to be zero because most of the PV power factor are unity.  $QD_i(h)$  is set by the given power factor. Thus, (1)-(12) become a stochastic optimal power flow problem.  $PG_i(h)$  and  $PD_i(h)$  are known random variables.  $E_{BESS}(n, h)$ ,  $P_{BESS}^{dis}(n, h)$  and  $P_{BESS}^{ch}(n, h)$  are control variables while others are state variables.

The above problem is solved by two loops: the outer loop is implemented by point-estimation method [15, 16] that considers both location coefficients and concentrations of random variables (PV and load) while the inner loop deals with the 24 h optimal power flow considering



deterministic PV generations and loads using the interior-point method [17].

Let the expected value, standard deviation and skewness coefficient of each random variable  $RD_k$  (that's,  $PG_i(h)$  and  $PD_i(h)$ ) be  $\mu_{rdk}$ ,  $\sigma_{rdk}$  and  $\lambda_{rdk}$ , respectively, where  $k=1, 2, \dots, 24 \times (Npv+Nbus)$  where  $Npv$  is the number of PV array. Let  $24 \times (Npv+Nbus)$  be identical to  $Nrd$ .

Step C1: Input the studied data.

Step C2: Compute the two perturbations:

$$\xi_{k,1} = \frac{\lambda_{rdk}}{2} + \sqrt{Nrd + \left(\frac{\lambda_{rdk}}{2}\right)^2}, \quad (13)$$

$$\xi_{k,2} = \frac{\lambda_{rdk}}{2} - \sqrt{Nrd + \left(\frac{\lambda_{rdk}}{2}\right)^2}, \quad k = 1, 2, \dots, Nrd \quad (14)$$

Also, compute the two weighting factors:

$$\omega_{k,1} = -\frac{\xi_{k,2}}{Nrd(\xi_{k,1} - \xi_{k,2})} \quad (15)$$

$$\omega_{k,2} = \frac{\xi_{k,1}}{Nrd(\xi_{k,1} - \xi_{k,2})} \quad k = 1, 2, \dots, Nrd \quad (16)$$

$$\sum_{k=1}^{Nrd} \omega_{k,1} + \omega_{k,2} = 1 \quad (17)$$

Step C3: Estimate the two location parameters:

$$RD_{k,m} = \mu_{rdk} + \xi_{k,m} \cdot \sigma_{rdk}, \quad (18)$$

$$m = 1, 2; k = 1, 2, \dots, Nrd$$

Step C4: Let  $k$  be 1.

Step C5: Let  $m$  be 1.

Step C6: Find the optimal  $E_{BESS}(n, h)$ ,  $P_{BESS}^{dis}(n, h)$  and  $P_{BESS}^{ch}(n, h)$  by the interior point method using all expected values of random variables except for  $RD_{k,m}$ .

Step C7:  $m=m+1$ . If  $m=2$ , then go to Step C6; else go to Step C8.

Step C8:  $k=k+1$ . If  $k > Nrd$ , then go to Step C9; else go to Step C5.

Step C9: Compute the final result: the mean and standard deviation of  $E_{BESS}(n, h)$ ,  $P_{BESS}^{dis}(n, h)$ ,  $P_{BESS}^{ch}(n, h)$ ,  $V_i(h)$  and  $CE_{loss}(h)$  using  $2Nrd$  sets of deterministic solutions.

Solving the problem in Step C6 becomes easy because it is a quadratic programming problem. This work employs the interior point method, which can obtain a local optimum, in MATLAB to solve this problem [17].

#### IV. SIMULATION RESULTS

A 33-bus distribution system [18] was used to study the problem using the proposed method, as shown in Fig. 1. One-MW PV array is at bus 18. The energy storage systems locate at buses 16, 17, 21, 22, 25 and 33.  $C_{peak} = 27$  \$/MWh and  $C_{off-peak} = 18$  \$/MWh. During 8:00 AM-16:00 PM, the BESS is discharging and  $C_{peak}$  is applied in (2). When it is 17:00 PM-7:00 AM, the BESS is charging and the off-peak tariff ( $C_{off-peak}$ ) is applied. The efficiencies of the BESS charging and discharging are 85% and 100%, respectively. Figs. 2 and 3 illustrate the total system loads in

24 hours and the individual bus loads at 7:00 AM, respectively. Figure 4 shows the 24-h solar irradiations.

##### A. Deterministic Results

First, the deterministic study was conducted by considering the deterministic irradiations and bus loads modeled by their corresponding mean values only. Figure 5 shows the loss profile in 24 hours. It could be found that the system has peak loads at 12:00, 14:00 and 15:00 PM, as shown in Fig. 2; however, the losses at these three hours are not the largest ones, compared to those in the evening. It is interesting to find that the largest loss occurs at 17:00 and 22:00 PM because there is no photovoltaic power delivered reversely to the substation at bus 1 in the evening.

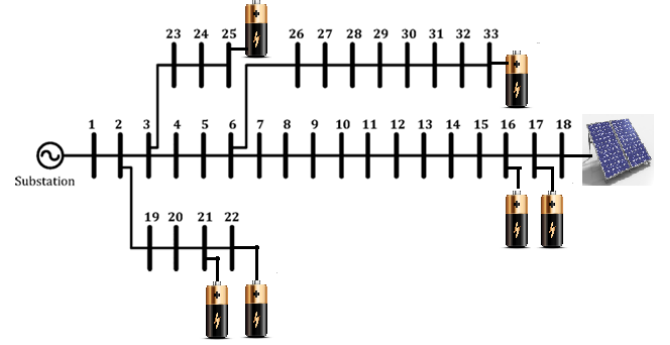


Fig. 1. One-line diagram of 33-bus system.

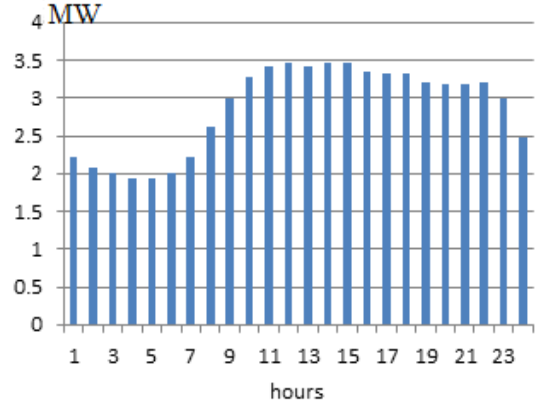


Fig. 2. Total system loads (MW) in 24 hours.

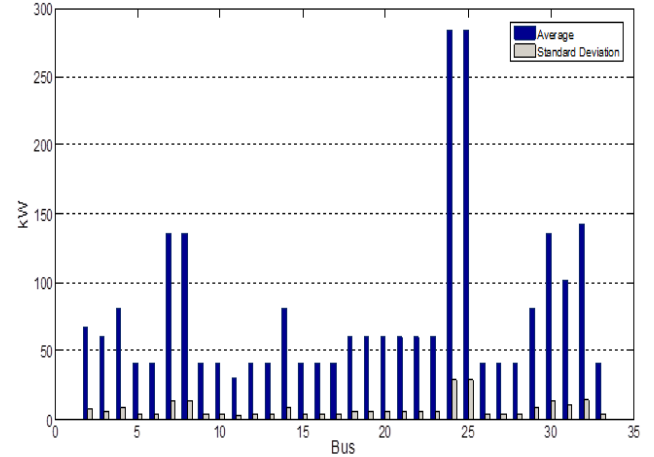


Fig. 3. The average values and standard deviations of kW demands at all buses at 7:00 AM.

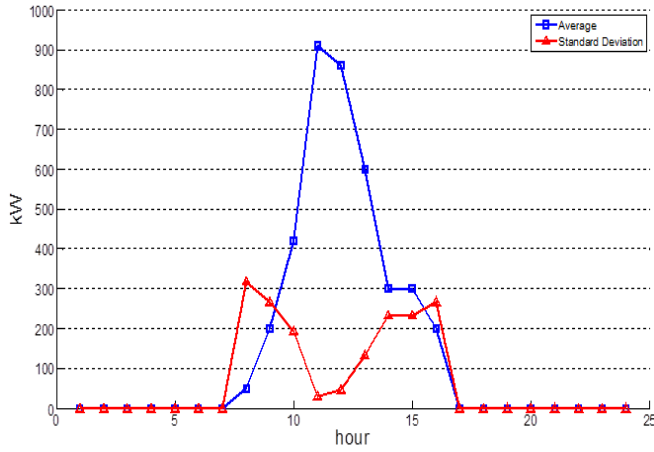


Fig. 4. The average value and standard deviation of photovoltaic power (kW) in a day.

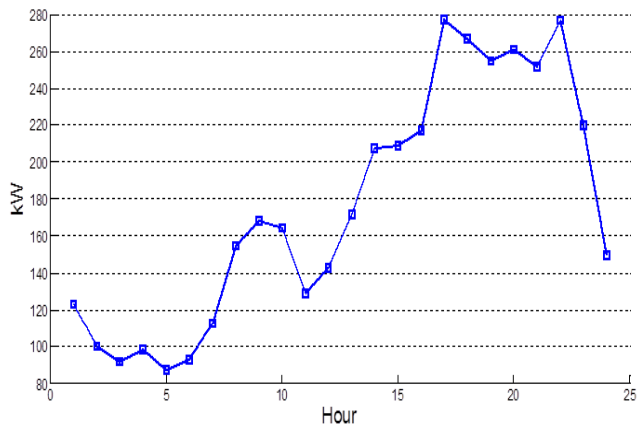


Fig. 5. Loss (kW) profile in 24 hours.

### B. Stochastic Results

This subsection considers the mean values and standard deviations of bus loads and solar photovoltaic power. Figure 6 shows the mean values of charging/discharging power (kW) for each BESS at those six buses. The positive (negative) values signify charging (discharging) of the BESS. The patterns of BESS at buses 21, 22, 25 and 33 are similar; however, those of the BESS at buses 16 and 17 are different because the BESS at buses 16 and 17 are near the photovoltaic array at bus 18. Figure 7 shows the standard deviation of charging/discharging power (kW) for each BESS at those six buses. It can be found the profiles of mean values are similar to those of standard deviations.

Figure 8 gives the mean value of energy (kWh) for each BESS while Figure 9 shows its corresponding standard deviation. It can be found the BESS is discharged during 8:00 AM-16:00 PM because  $C_{peak}$  is high and discharging energy can reduce the cost. When it is 17:00 PM-7:00 AM, the BESS is charged because cheap kWh from the utility can be stored in the BESS. The standard deviations of kWh have a local minimum near 12:00 because the standard deviation of irradiation also is near zero at 12:00, as shown in Fig. 4. Essentially, the power (energy) charged to the BESS at buses 21, 22, 25, 33 are from the substation while that charged to the BESS at buses 16 and 17 is mainly from the PV array. The energies at buses 16 and 17 are discharged moderately during 8:00 AM-16:00 PM while those at other buses are discharged linearly.

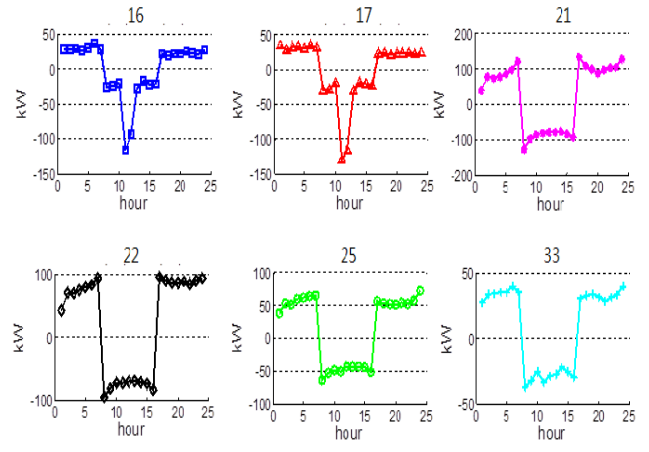


Fig. 6. The mean value of charging/discharging power (kW) for each BESS at six buses.

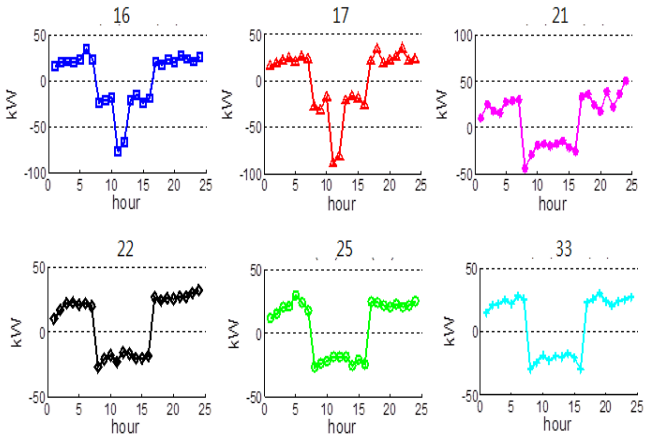


Fig. 7. The standard deviation of charging/discharging power (kW) for each BESS at six buses.

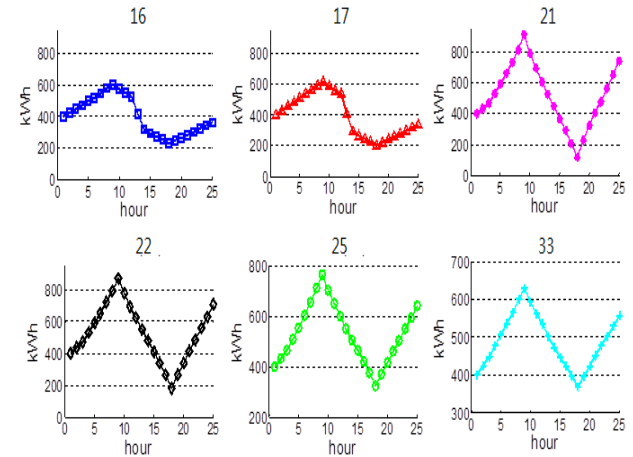


Fig. 8. The mean value of energy (kWh) for each BESS.

Figures 10 and 11 show the profile for the mean values and standard deviations of kW losses in 24 h, respectively. The mean value and standard deviation of loss cost ( $CE_{loss}$ ) are \$33.38 and \$16.03, respectively. The deterministic cost is \$90.16, which is greater than the estimated upper bound ( $81.47=33.38+3\times16.03$ ). Thus, the deterministic loss cost provides a pessimistic (conservative) solution while the stochastic loss cost is an optimistic solution with uncertainty. The largest loss occurs at 17:00 that is the same as the time occurring in the deterministic case.

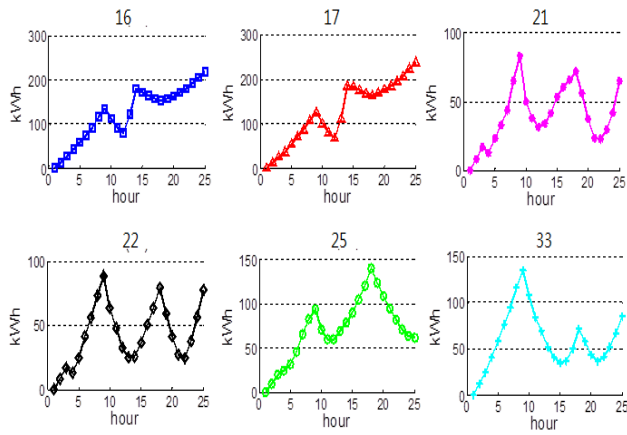


Fig. 9. The standard deviation of energy (kWh) for each BESS.

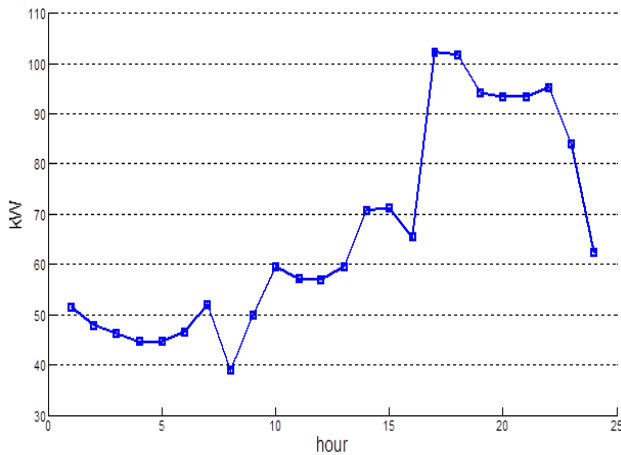


Fig. 10 The profile for the mean values of kW losses in 24 h.

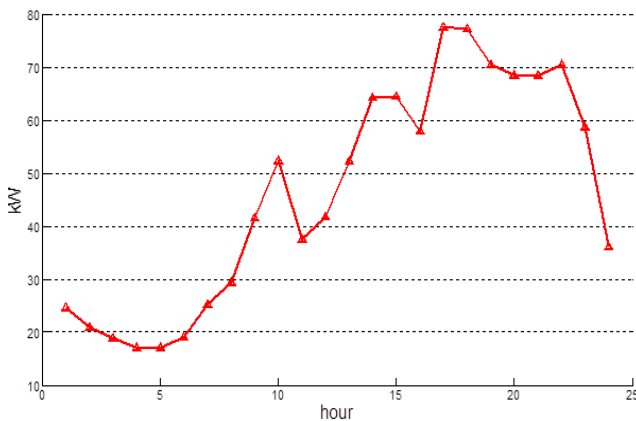


Fig. 11. The profile for the standard deviation of kW losses in 24 h.

## V. CONCLUSIONS

In this paper, a method based on a two-loop algorithm is proposed to study the day-ahead energy scheduling of battery energy storage system in a distribution system. The outer loop deals with the random variables (PV and load) while the inner loop deals with the 24 h optimal power flow. There are some key findings as follows.

- (1) The largest loss occurs at nightfall because the PV power generation starts to decrease.
- (2) The BESS can help the grid regulate the power generation from both the PV and substation,

depending on the locations of BESS, to reduce the loss cost.

- (3) The stochastic solution provides an optimistic result with mean value and standard deviation.

## REFERENCES

- [1] A. Poullikkas, A comparative overview of large-scale battery systems for electricity storage, *Renewable and Sustainable Energy Reviews*, Renewable and Sustainable Energy Reviews, vol. 27, pp. 778–788, 2013.
- [2] A. Keyhani, *Design of Smart Power Grid Renewable Energy Systems*, Wiley, 2011.
- [3] J. Quesada, R. Sebastián, M. Castro, J.A. Sainz, Control of inverters in a low-voltage microgrid with distributed battery energy storage. Part II: Secondary control, *Electric Power Systems Research*, vol. 114, pp. 136–145, 2014.
- [4] I. Serban, C. Marinescu, Battery energy storage system for frequency support in microgrids and with enhanced control features for uninterruptible supply of local loads, *Electrical Power and Energy Systems*, vol. 54, pp. 432–441, 2014.
- [5] M.Z. Daud, A. Mohamed, M.A. Hannan, An improved control method of battery energy storage system for hourly dispatch of photovoltaic power sources, *Energy Conversion and Management*, vol. 73, pp. 256–270, 2013.
- [6] S. Koohi-Kamali, N.A. Rahim, H. Mokhlis, Smart power management algorithm in microgrid consisting of photovoltaic, diesel, and battery storage plants considering variations in sunlight, temperature, and load, *Energy Conversion and Management*, vol. 84, pp. 562–582, 2014.
- [7] M. Khalid, A.V. Savkin, Minimization and control of battery energy storage for wind power smoothing: Aggregated, distributed and semi-distributed storage, *Renewable Energy*, vol. 64, pp. 105–112, 2014.
- [8] D. Mehdi, B. Jamel, R. Xavier, Hybrid solar-wind system with battery storage operating in grid-connected and standalone mode: Control and energy management- Experimental investigation, *Energy*, vol. 35, pp. 2587–2595, 2010.
- [9] I. Duggal and B. Venkatesh, “Short-term scheduling of thermal generators and battery storage with depth of discharge-based cost model,” *IEEE Trans. on Power Systems*, vol. 30, no. 4, pp. 2110 – 2118, 2015.
- [10] M. Kahl, C. Freye, T. Leibfried, “A cooperative multi-area optimization with renewable generation and storage devices,” *IEEE Trans. on Power Systems*, vol. 30, no. 5, pp. 2386 – 2395, 2015.
- [11] A.C. Luna, N.L. Diaz, M. Graells, J.C. Vasquez, J. M. Guerrero, “Mixed integer linear programming-based energy management system for hybrid PV-wind-battery microgrids: modeling, design, and experimental verification,” *IEEE Trans. on Power Electronics*, vol. 32, no. 4, pp. 2769 – 2783, 2017.
- [12] H. Wu, M. Shahidehpour, A. Alabdulwahab, A. Abusorrah, “Thermal generation flexibility with ramping costs and hourly demand response in stochastic security-constrained scheduling of variable energy sources,” *IEEE Trans. on Power Systems*, vol. 30, no. 6, pp. 2955 – 2964, 2015.
- [13] Y. Degeilh, G. Gross, “Stochastic simulation of utility-scale storage resources in power systems with integrated renewable resources,” *IEEE Trans. on Power Systems*, vol. 30, no. 3, pp. 1424 – 1434, 2015.
- [14] R.A. Jabr, S. Karaki, J.A. Korbane, “Robust multi-period OPF with storage and renewable,” *IEEE Trans. on Power Systems*, vol. 30, no. 5, pp. 2790 – 2799, 2015.
- [15] A.R. Malekpour and T. Niknam, “A probabilistic multi-objective daily Volt/Var control at distribution networks including renewable energy sources,” *Energy*, vol. 36, pp. 3477–3488, 2011.
- [16] M. Aien, M. Fotuhi-Firuzabad, M. Rashidinejad, “Probabilistic optimal power flow in correlated hybrid wind-photovoltaic power systems,” *IEEE Trans. on Smart Grid*, vol. 5, no. 1, pp. 130 – 138, Jan. 2014.
- [17] Optimization Toolbox- fmincon, MATLAB, The MathWorks, 2009.
- [18] B. Venkatesh, R. Ranjan, and H. B. Gooi, “Optimal reconfiguration of radial distribution systems to maximize loadability,” *IEEE Trans. Power Systems*, vol. 19, no. 1, pp. 260–266, Feb. 2004.

# Improving the Adaptivities of Over-The-Top Television System

Ha Tran Thu

*Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education( HCMUTE)  
Ho Chi Minh City, Vietnam  
thuha@hcmute.edu.vn*

Son Tran Minh

*HUTECH Institute of Engineering,  
Ho Chi Minh City University of Technology  
(HUTECH)  
Ho Chi Minh City, Vietnam  
soq2000@yahoo.com*

Thai Nguyen Van

*Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education ( HCMUTE)  
Ho Chi Minh City, Vietnam  
thainv@hcmute.edu.vn*

Minh Le Hoang

*Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education( HCMUTE)  
Ho Chi Minh City, Vietnam  
minhhlh@hcmute.edu.vn*

**Abstract**—Over The Top technology (OTT) – thanks to the adaptive streaming, *i.e.* automatically adjusting the video stream's quality – quickly takes over the role of its precedent IPTV and becomes the indispensable service for any provider of streaming video over the Internet. However in deployment, there are still several white spaces to improve the technology. In this paper we propose 3 techniques toward that goal. Firstly, taking into account the limited resource of the encoders, we propose a temporal and spatial determination of the pre-generated bitstreams based on the statistical analysis of network traffic. The number of streams remains unchanged but they fit better to the real condition of the transmission network. Secondly, addressing the booming number of available digital TV programs, we provide an adaptively customized catalogue of the TV programs. The broadcasted programs are organized by their contents' natures. Viewers' historical interaction with TV is incorporated to further fine-tuning the program lists proposed to viewers themselves. Finally, facing to the extreme situation when the OTT servers can still reach to the saturated state, we propose to switch adaptively between serve-client and peer-to-peer topologies to retrieve the requested TV program in an efficient way. These three adaptivity-methods are combined together in a full chain of OTT distribution system to be evaluated on their overall performance. The impact is expected to be pertinent.

**Keywords:** *Over The Top technology (OTT); the Internet Protocol Television (IPTV); peer-to-peer (P2P).*

## I. INTRODUCTION

OTT technology – the descendant of the IPTV – quickly takes over the role of its precedent and becomes the indispensable service for any provider of streaming video over the Internet. Thanks to the adaptive streaming – automatically adjusting the video stream's quality – the OTT content can resist against the bandwidth fluctuation due to the uncertainty of the open Internet, the situation where the conventional IPTV programs never expose their contents. Hence, IPTV providers choose the safer mode of delivery. They set up their own Internet infrastructure and conduct physically lines to each of their subscribers to ensure the quality of the service. OTT service providers nowadays can truly offers to larger

audiences outside of their infrastructure (or without investing at all the transmission infrastructure) at the considerably reduced cost. However the OTT technology can be considered as technical solution dealing with the quickly varied bandwidth of the internet. In practice, there are still several white spaces to improve the technology.

3 techniques are analyzed to improve the adaptivity of the OTT technology:

1) A temporal and spatial determination of the pre-generated bitstreams based on the statistic analysis of the targeted regions to provide the service,

2) fine-tuning the program lists proposed to viewers to ease the channel selection,

3) selecting the closest viewers in order to reuse its content for saving bandwidth.

The above three adaptation-methods are combined together in a full chain of OTT distribution system to evaluate the overall performance. The impact is expected measurable.

The paper is organized as the following. The next section gives an overview on the television progress in the Internet era. Section 3 is dedicated to the detailed proposals on improving the OTT adaptivity. The paper is closed with conclusion and perspectives in the Section 4.

## II. OVERVIEW THE TELEVISION EVOLUTION IN THE INTERNET ERA

In the 21<sup>st</sup> century the access with broadband internet and downstream data rates of several Megabit per second (Mbit/s) is making a steady progress. With the increasing number of households having access to Internet, using Internet in every daily life for both personal and professional tasks, Internet brings a disruptive changes to the Television Industry (see [1] and [2]). According to [3] we are currently in the third wave of the television technology in the Internet Era. The first wave, covering the years 2004-2008, was characterized by user-generated web videos, applied primarily as a streaming video model, and included early video initiatives from network television broadcasters (Fig. 1). The second wave, from 2009 - 2011, saw the introduction of a new class of



participants, including but not limited to Netflix, Amazon, and other retailers. In this wave, traditional television broadcasters still play a dominant role while enlarging the television services, offering additional device accessibility to support the paid television model. The third (being the current wave) cannot be easily characterized because of its rapidly evolving and transforming nature. Thus far, this wave has introduced device-to-television beaming, a customer choice of ad-free or ad-supported content, and significantly, the offering of OTT service delivery platforms from a variety of organizations. These organizations do not necessarily possess costly infrastructure for television transmission as traditional providers often have.

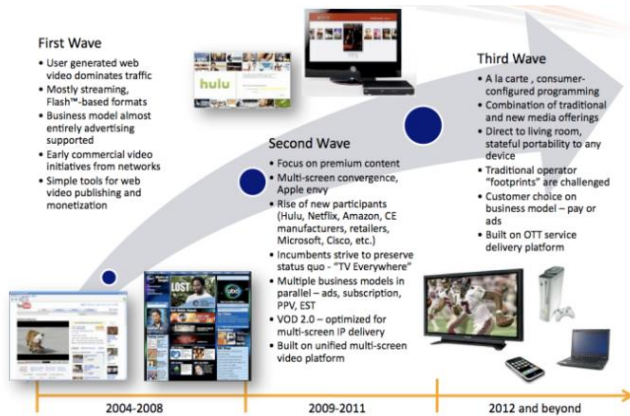


Fig. 1. Three waves of the Television technology in the Internet Era

Around the world, OTT technology has changed the face of classic digital television. Multimedia content (playback or live view) can be no longer exclusively distributed by media giants equipped with expensive cable and satellite infrastructure. Television content owned by small companies that cannot afford to develop their own network can still be supplier to viewers via an open Internet of acceptable quality [4].

Fig. 2 shows up some of the traditional giant television channels in US such as Comcast, Dish (cyan paths in Fig. 2). For instance Dish in 2017 possessed 30M subscribers via their 10 satellites networks. Fig. 2 also presents some purely content providers like Hulu and Netflix. They are typical model for the application of OTT in transmitting contents to end-users. The success of these models (104 million subscribers for Netflix possessing zero media infrastructure) has enforced the classical television providers to rethink their business model to be compatible with the tendency of the Internet era.

Fig. 3 outlines the architecture of IPTV together with the conventional infrastructure of digital television broadcaster (see [5]). It is a typical architecture for service provider during the first and the second wave of the television (Fig. 1). With IPTV, the managed IP network – setup and owned by the service provider – is the new transmission platform for distributing television programs besides the conventional ones such as terrestrial, cable and satellite networks.

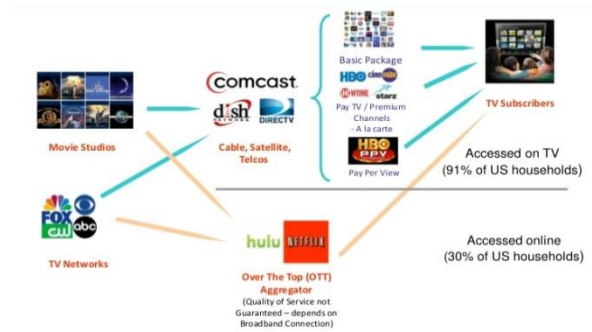


Fig. 2. Some US stakeholders in the television industry in the Internet Era

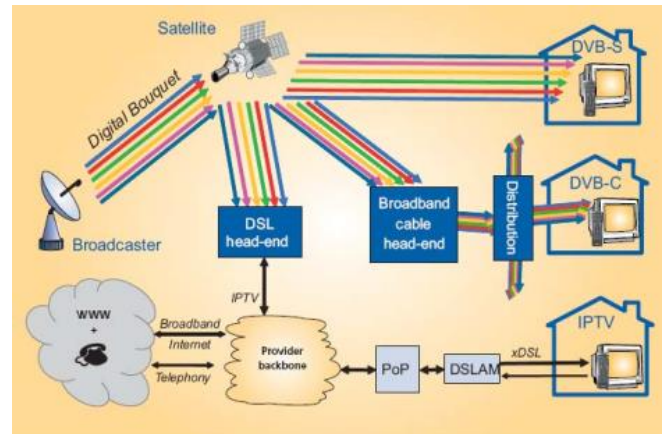


Fig. 3. Main architecture of a Digital Video Broadcasting with IPTV service.

In the third wave of the television technology, the OTT service providers deliver audio, video and other media over the internet and bypass the traditional operator's network. Since, the OTT players do not require any business or technology affiliations with network operators for providing such services, they are often known by the term "Over-The-Top" (OTT) applications [6]. Fig. 4 demonstrates the simple network architecture of an OTT service.

In Vietnam, several OTT services – supplemental services of the traditional television channels or Internet providers – can be taken into account such as FPTPlay, SCTVOnline, VTVGo (see [7] and [8]). ClipTV for instance is a pure OTT provider in Vietnam.

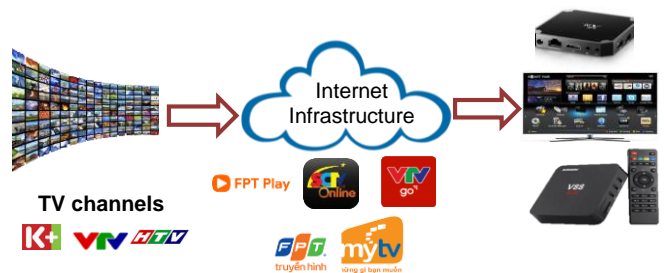


Fig. 4. Open Internet is considered as the distributing infrastructure of OTT service providers.

### III. IMPROVING POSSIBILITIES FOR OTT

The new OTT technology, rooted deeply on the Internet, has made distribution of audiovisual content less complicated in the third wave of the television industry. The often-repeated mantra that "while content is king" it is "distribution that's the emperor", suggesting that strongly-capitalized operators investing on their own cable networks (or other transmission network as in Fig. 3) are the winners in the television industry,



is starting to lose its relevance as the open internet gradually starts to replace traditional managed distribution for many consumers [11].

However, the OTT is not yet a perfect / superior solution compared to IPTV. In [3], one can find a lot of challenges for the technology that the correct answers are still opened. In the scope of this paper, we address 3 following issues:

1) The challenge of image quality: "Blocking effect" whenever the open Internet's infrastructure is degraded in data bandwidth.

2) The list of OTT channels / content is not appropriately adjusted for each audience. The explosion in the number of channels and multimedia contents makes it increasingly time-consuming for viewers to choose the right content to watch.

3) The "Bottleneck" effect happens whenever there are number of viewers accessing at the same time the OTT service.

In the following, we will focus on the possible solutions for the above problems.

#### A. Adaptive streaming for better fit to the channel condition

Although broadband internet and downstream data rates of several Megabit per second (Mbit/s) is making a steady progress nowadays, the open Internet provided by a third-party is likely unstable for sensible and also bit-consuming video transmission. It is the reason that during the first two waves of Television industry, the IPTV technology was based firmly on the managed IP network. In other words, only the closed Internet – private IP infrastructure owned by the television service itself – can manage to exclusively allocate enough bandwidth to video transmission. With the higher and higher speed Internet available to public, the room left for television service is not saturated quickly and frequently. It gives an opportunity for OTT to thrive thanks to a key technology called adaptive streaming. A television program is split into several segments, each segment is then cloned several times for the same carried content but at various bitrates (hence various qualities). Whenever there is a fluctuation in the Internet transmission, the OTT television can carefully select the cloned segment having the best bitrate still passing through the available bandwidth. Thanks to this virtue, the television program can be viewed in continuity and with an acceptable quality.

It is evident that if one could clone infinitely each segment for all possible bitrates, the OTT service would be the best optimal facing to the instability of the open Internet. In practice, the generation of infinite number of bitrates for the same content – especially for the real-time requirement for live televisions – is almost impossible from the viewpoint of resource consumption and economy. That is why it is recommended to have 4 bitrates for each segment in any situation of the Internet.

Fig. 5 resumes how Internet speeds fluctuate throughout the day within different types of conurbation in England [9]. It can be seen that the average of available bandwidth changes according to geographical and temporal factors. For the practical reason, if only 4 static bitrates are fixed to provide the television service for all regions at all times, the OTT service will cause a bad image quality with a lot of blocking effect. It is because a region equipped with faster network or the interval with better network-condition can never benefit

from their advantages to warrant the acceptable quality for the worse region, worse period.

Hence, as the first improved adaptivity method, we propose to perform a comprehensive analysis of the channel capacity throughout spatial and temporal factor. The deduced most probable bandwidth ranges are then taken into account to determine the best fitted N values (N=4 for most of the case) of bitrates to clone / re-encode each segment of television program.

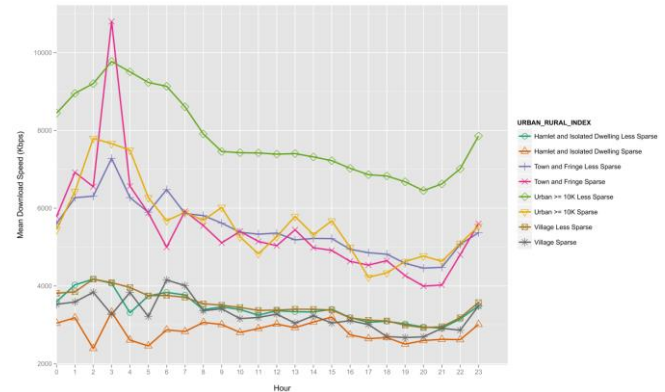


Fig. 5. A temporal fluctuation in bandwidth of the open Internet.

#### B. Adaptive programming for better meeting audience's demand

Second improvement proposed in this paper is to enhance the experience of tele-spectators in selecting / viewing channels. The emergence of digital television brings to viewers not only quality but also quantity of the television channels. It is not rare that with several carrier frequencies (correspond to only several analogous televisions) hundreds of channels can be tuned in (Fig. 6). Choosing a right channel at a right moment becomes a very hard work – if not a nightmare - for tele-spectators.

A recommended program adaptively presented to viewer can be a perspective direction to tackle the above problem. Thanks to the Event Information Table embedded in the digital video stream, the theme of the program at the given moment can be collected and analyzed. Then tele-spectators will be offered with a program adaptively recommended according to the interest and habitude. For instance, during the dinner time, the habitual channel will be tuned in; a list of interested topics will be shown up ordered by theme, transparently with the underlying physical digital channel.



Fig. 6. Abundant and therefore chaotic offer of digital television channels

### C. Adaptive selecting source for efficient usage of transmission capacity

The acceptable quality of OTT service over the open Internet is the result of the marriage between adaptive streaming (Section A) and Content Delivery Network (CDN). While OTT players are the front end for the consumer, CDNs are the essential backend responsible for delivering the television appropriately cloned segments. CDNs are proxy web servers that deliver content to end consumer based on the proximity to the end user. However the CDN capacity of continuous delivery is not unlimited. Due to the server-client topology, the CDNs are also faced to the saturated condition whenever there are huge numbers of tele-spectators.

Fig. 7 outlines the peak-evolution of the required bandwidths for several exceptional events attracting a lot of television viewers [10]. Obviously, for the economy reason, OTT service providers do not plan to cover such extreme situation. Even worse, the capacity of the underlying CDNs is already drained out at the prime-time every weekend.

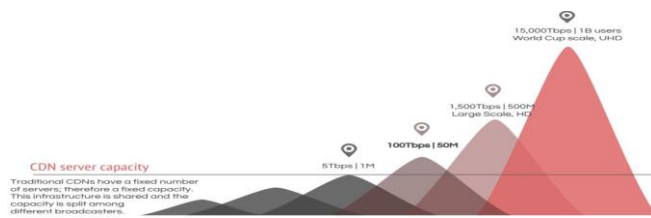


Fig. 7. Limitation of CDNs in transmitting OTT television programs

In order to rise the OTT service to a higher grade of being independent from the network infrastructure, we propose a hybrid transmission of the content from the original servers to the OTT players located at the tele-spectators' side. In order to view an OTT program, the sources of the program are adaptively selected to overcome the limit of the CDNs. The nature of television is the live program and broadcasted in mass for coincident viewing. Therefore reusing the source from a peer – a neighbor tele-spectator having closer and faster connection than from the original server itself – can flatten the peak dilemma. The content distribution in this case becomes peer-to-peer (P2P) topology rather than the server-client one ([12]). Furthermore, adaptively selecting source via P2P can partially take over the role of CDNs, a paid service often provided by a third party. The cost for CDNs is somehow considered as the monthly investment in managed infrastructure in the case of IPTV, which is just contradictory to the OTT virtue of infrastructure-free.

We propose to request simultaneously the same segments from neighbor peers and from the original server. Whenever the server is saturated, the OTT program can be still viewed smoothly thanks to the provision from the peers. To really save the server's bandwidth, a kind of weighed Interactive Connectivity Establishment is incorporated. A cost metrics will be permanently calculated and maintained based on the geographical locations and the estimated bandwidth between the peers (including the requesting OTT player and the potentially serving peers). Evaluation of this metrics in real-time can deduce the best candidate to provide the OTT segments. If the original server is not selected, the request will be temporarily redirected to the determined candidate.

### IV. CONCLUSION AND PERSPECTIVES

Nowadays we are the witness of the third wave of the television industry, whereby the OTT becomes more and more

widely-spread in the television industry. The key-difference is mainly attributable to OTT distribution being separated from traditional distribution infrastructure and “runs on an open internet connection without the benefit of a managed network” [11]. Its infrastructure-separated nature allows for disruptive innovations in the value chain but also causes critical issue that makes OTT substantially different to other historical technological advancements in the television industry.

The paper proposes to fine-tune the adaptivity of the OTT technology – the key-success of OTT over the open Internet. Three directions of adjusting the adaptivity are possible. Firstly the adaptive streaming scheme can be dynamically adjusted to fit the spatial and temporal condition of the transmission network. Secondly the abundant choice of OTT programs can be adaptively present to viewers according to their habitude and interested themes. Thirdly, an adaptive selection of the programs' source is proposed to combine the server-client and peer-to-peer topology in order to improve the bandwidth saturation of the OTT service, to reinforce the virtue of the infrastructure-free architecture.

The intention of the work is to contribute to the research and development of the OTT technology, which is at heart of the third wave of the television industry. A disruptive model of transmission – infrastructure independent – requires exceptional efforts in all aspects to reach to the OTT goal: enriching the video content by democratizing the infrastructure usage.

### REFERENCES

- [1] J. C. Whitaker, *Interactive TV Demystified*, 2001 McFraw-Hill, ISBN 0-07-136325-4
- [2] E. Diehl, *Securing Digital Video*, 2012 Springer, ISBN 978-3-642-17344-8.
- [3] T. Ohanian, *Over-the-Top Considerations: Functionalities and Technologies*, Cisco Systems, NAB 2014.
- [4] Y. N. K. Chen, *Competitions between OTT TV platforms and traditional television in Taiwan: A Niche analysis*, 2018 Elsevier Ltd.
- [5] A. Punchihewa, *Tutorial on IPTV and its latest developments*, ICIAFS January 2011.
- [6] L. Bringuier, *White Paper OTT Streaming – 2<sup>nd</sup> edition*, September 2011, Anevia
- [7] Asia pacific Pay-TV distribution, *The future of Pay-TV & Fixed Broadband in Asia*, 2018 Media Partners Asia
- [8] Asia pacific Online Video & Broadband distribution, 2018 Media Partners Asia
- [9] D. Riddlesden, A. D. Singleton, *Broadband speed equity: A new digital divide?*, 2014, Applied Geography, Elsevier.
- [10] Official Website of StreamRoot, <https://streamroot.io>.
- [11] C. Waldenor, *Is OTT Disrupting Television?* Master Thesis, Stockholm, June 7<sup>th</sup> 2013.
- [12] E. Setton, B. Girod, *Peer-to-Peer Video Streaming*, 2007 Springer Science+Business Media, LLC, ISBN-13: 978-0-387-74114-7.

# Multi-class Support Vector Machine Algorithm for Heart Disease Classification

Thanh-Nghia Nguyen

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, VietNam  
nghiant@hcmute.edu.vn

Thanh-Hai Nguyen

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, VietNam  
nthai@hcmute.edu.vn

Duc-Dung Vo

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, VietNam  
dungvd@hcmute.edu.vn

Truong-Duy Nguyen

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, VietNam  
duynt@hcmute.edu.vn

**Abstract**—It is very necessary to build a tool for critical diagnosis of heart disease related to human health. In this paper, a state-of-art method with multi-class support vector machine is proposed to classify types of heart disease in different scenarios of heartbeat datasets. In practice, electrocardiogram signals are preprocessed using a median filter to obtain smooth electrocardiogram signals and then the smooth electrocardiogram signals are segmented to obtain heartbeats. A short-time Fourier transform algorithm is applied on heartbeats to produce frequency information, which can be seen as features of heart disease. Therefore, the features are the input of the multi-class support vector machine classifier for training and testing. In addition, the electrocardiogram signals are collected from the MIT-BIH database for this research, in which there are five types of heart disease with two cases of intra- and inter-patient. The experimental results show that the high accuracy is obtained to illustrate the effectiveness of the proposed classification system.

**Keywords**— Short-time Fourier Transform; Electrocardiogram signal; Multi-class Support Vector Machine; heart disease classification

## I. INTRODUCTION

An electrocardiogram (ECG) signal has small waves such as P, Q, R, S, and T which can contain features related to heart disease. ECG signals are often collected from ECG machines through electrodes, which are installed on the surface of the human body. In addition, an ECG signal can be observed on a screen or on ECG papers for diagnosis. Therefore, doctors can determine exactly heart disease based on the ECG signal. Moreover, ECG signals can be employed to study heart disease by scientists through processing and analyzing them for determining different types of heart disease.

Noise reduction in ECG signals plays an important role to obtain the pure ECG signals. Noise and artifact in the ECG signals include power-line interference, motion artifacts, baseline wander, muscle noise, and other types of interference. Therefore, many different methods, which have been developed in recent years to cancel unwanted noise components in the ECG signal, are applied such as Finite Impulse Response (FIR) filter, Infinite Impulse Response

(IIR) filter, Grey Spectral Noise Cancellation (GSNC) and Wavelet Transform with thresholds [1-5]. In particular, Yang Xu *et al.* [6] proposed the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and a wavelet threshold for the correction of the baseline wander noise. With power-line interference noise, Manpreet Kaur Aneja *et al.* [7] proposed digital notch filters methods for canceling the power-line interference noise.

Characteristics of ECG signals are shaped with the basis of the waveform including P, Q, R, S, and T waves [8]. From the basis waveform, it creates some of features such as QRS interval, R-R interval, P-R interval, S-T interval, Q-T interval, P-R segment, and S-T segment. The heart disease can be diagnosed based on the features related to the waves. Therefore, the best features of heart disease from the ECG signal can achieve more accurately using classifiers. There are a lot of feature extraction methods have been used such as wavelet transform, z transform, adaptive threshold and Principal Component Analysis (PCA), normalized RR intervals and morphological features, combination between features in the time and frequency domains using Random Forests, the Linear Discriminant Analysis (LDA) and the Independent Component Analysis (ICA) [9-13].

Short-Time Fourier Transform (STFT) algorithm is a Fourier-related transform and the procedure for computing the STFT is to divide a longer time signal into shorter segments of equal length. From the shorter segment of the signal, it is computed using the Fourier transform separately on each shorter segment. Therefore, frequency features of the signals corresponding to each segment are obtained using the STFT algorithm. In recent years, the STFT algorithm has been applied on ECG signals to extract the features [14-16]. In particular, the STFT algorithm can be combined with chaos analysis to detect the QRS complex. In addition, this method can be employed to convert the ECG signal to an image, which is the feature of heart disease for training and testing based on Convolution Neural Networks (CNNs) for classifying the heart disease.

Classification of heart disease based on ECG signals is an important task to understand heart condition of human.

Moreover, the classification and detection of abnormal heartbeat can help doctor and patient to know about the heart disease for treatment soon. In recent decades, a lot of methods have been utilized for classifying heart disease with high accuracy including kernel PCA and Neural Network (NN), Generalized FFNN, Modular neural network, Feed forward Probabilistic Neural Network (PNN), Support Vector Machine (SVM) classifier with Kernel-Adatron (KA), Cascade forward Back Propagation Neural Network (BPNN), and CNN [14, 17, 18].

This paper is organized as follows: Section-2 shows the method of STFT for obtaining frequency information and the multi-class SVM for classifying heart disease. The Section-3 presents simulation results and discussions. The final section describes the conclusions of the article.

## II. MATERIAL AND METHOD

### A. Proposed Method

In this paper, the proposed classification method for classifying five types of heart disease is a Multi-class Support Vector Machine (MSVM), in which the classification method consists of six stages including ECG signals, pre-processing, heartbeat segmentation, feature extraction, multi-class SVM classifier and results of heart disease classification as shown in Fig.1. In particular, the ECG signals downloaded from the MIT-BIH database are filtered by a median filter to smooth the ECG signals. In the third stage, the smoothing ECG signals are segmented into heartbeats in the time domain and then the STFT algorithm is applied to transform the heartbeats to the frequency domain considered as features. Therefore, the features of heartbeats are used in the multi-class SVM classifier for classifying heart disease and then results are evaluated based on a confusion matrix method.

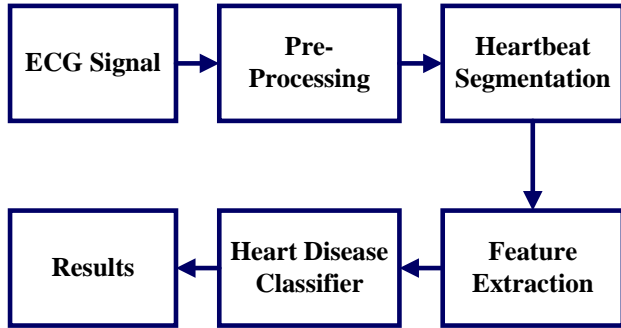


Fig. 1. Block diagram of the proposed classification system

### B. Short-Time Fourier Transform for Feature Extraction

ECG signals after filtering are segmented to produce heartbeat signals. The STFT algorithm with a window function is applied to extract frequency features of the heartbeat signals for classification. Therefore, the STFT is described as follows:

$$STFT\{x(n)\} \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x(n)d(n-m)e^{-j\omega n} \quad (1)$$

in which,  $x(n)$  presents for a heartbeat segmented from the smooth ECG signal,  $d(n)$  is a window function and  $X(m, \omega)$  denotes the Fourier Transform of the heartbeat in the window of  $d(n)$ .

In this paper, the rectangle window  $d(n)$  used in the STFT is described as follows:

$$d(n) = \begin{cases} 1, & 0 \leq n \leq M-1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

### C. Multi-class Support Vector Machine for heart disease classification

A MSVM with a supervised machine learning is employed in this research. A SVM algorithm is to determine the optimal hyperplanes between data points of different heart disease from input ECG data. Given a pair of training data  $\{x_i, y_i\}$ , the optimization issue of the SVM is defined as follows:

$$\min_{w, \zeta, b} J(w, \zeta) = \frac{1}{2} w^T w + C \sum_{i=1}^N \zeta_i \quad (3)$$

$$\text{subject to } y_i(w^T \varphi(x_i) + b) \geq 1 - \zeta_i, i = 1, \dots, N \quad (4)$$

$$\zeta_i \geq 0, i = 1, \dots, N \quad (5)$$

in which,  $w$  is the weight vector for training parameters,  $C$  is the positive regularization constant value, and  $\varphi$  is the mapping function used to map input data point  $x_i$  into a higher dimensional space. Moreover,  $\zeta_i$  is the positive slack variable which indicates the distance of  $x_i$  with respect to the decision boundary. To tackle the issue of the SVM, the Lagrange multipliers is applied to rewrite this expression as follows:

$$L(x) = \sum_{x_i \in SV} \alpha_i y_i K(x, x_i) + b \quad (6)$$

where,  $\alpha_i \geq 0$  is the Lagrange elements and  $K(x, x_i)$  is the kernel function of the SVM is defined as follows:

$$K(x_i, x_j) = \alpha_i (x_i)^T \alpha_j (x_j) \quad (7)$$

In this paper, an one-against-one (OAO) multi-class classifier is applied, so the kernel function  $K(x, x_i)$  with the Gaussian radial basis function is describes as follows:

$$K(x_i, x_j) = e^{-\sigma \|x_i - x_j\|^2} \quad (8)$$

## III. EXPERIMENTAL RESULTS

ECG signals, which are collected from the MIT-BIH arrhythmia database, consist of 44 recordings with the length of 30 minutes and the sampling frequency of each signal is 360 Hz. These ECG signals have five types of heart disease recommended by ANSI/AAMI EC57:2012 standard. In practice, the five types of heart disease are labeled with the letters of N, S, V, F, and Q. With 44 recordings from the MIT-

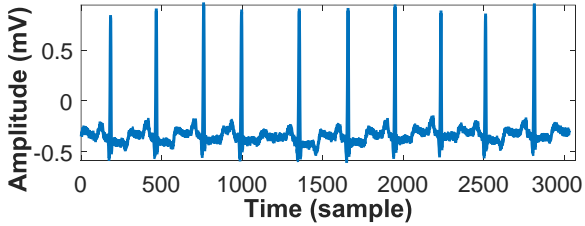


BIH database, we divided these ECG signals into two sub-datasets including DS1 = (101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230) and DS2 = (100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234). Moreover, all 44 recordings are called ADB including DS1 and DS2. Five types of heart disease related to three groups of DAB, DS1 and DS2 are described in Table I.

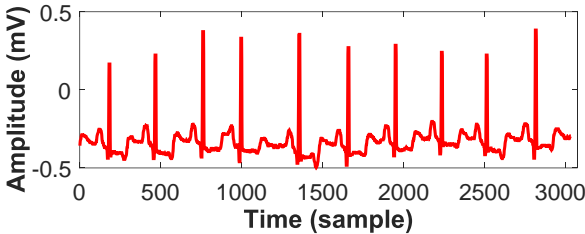
TABLE I. THE NUMBER OF HEARTBEATS WITH FIVE TYPES OF HEART DISEASE IN EACH DATASET

Dataset	Number of heart disease					
	N	S	V	F	Q	Total
DB1	45769	940	3782	414	1498	52403
DB2	44156	1834	3217	388	1461	51056
ADB	89925	2774	6999	802	2559	103459

To enhance the heart disease classification performance, each ECG signal was pre-processed using the 12<sup>th</sup>-order Median Filter (MF) to smooth. Moreover, the smoothing ECG signal was separated into heartbeats for extracting the heart disease feature. The result of applying the MF with the 12<sup>th</sup>-order is shown in Fig.2, in which the original ECG signal is the blue plot and the smoothing ECG signal is red.



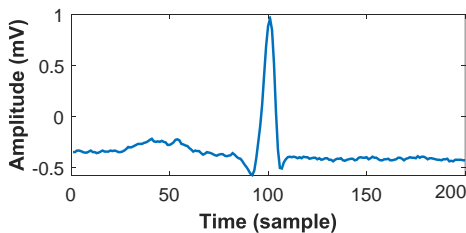
a) Original ECG signal



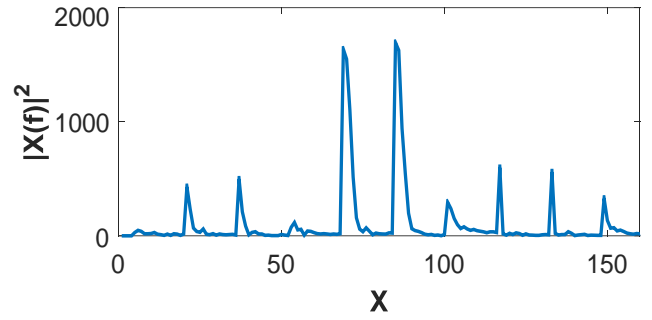
b) Smooth ECG signal applying median filter

Fig. 2. Representation of original ECG signal and smooth ECG signal

From heartbeats after separated from the smoothing ECG signal, the STFT was applied for the heartbeat to extract features. In this paper, the rectangle window with the size of 2<sup>5</sup> was used in the STFT algorithm for determining the short duration corresponding to frequency features extracted as shown in Fig.3. In particular, Fig. 3a is one heartbeat and heartbeat features are described in Fig. 3b. Therefore, the features are the input of the SVM classifier for training and testing to classify five types of heart disease.



a) One heartbeat signal



b) Heartbeat features after the STFT

Fig. 3. Representation of heartbeat and corresponding the STFT signal of the heartbeat.

To evaluate the heart disease classification performance, the evaluation metrics of Sensitivity ( $SEN$ ), Specificity ( $SPE$ ) and Accuracy ( $ACC$ ) as in (9-11) based on the confusion matrix was utilized. In particular, the definition of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) of confusion matrix is described as in Table II. In addition, the TP denotes correctly predicted positive class and the TN indicates correctly predicted negative class of heart disease. The FP is the instances of the other classes that are incorrectly classified to the given class and the FN is the instances from the given class that are incorrectly classified to another class. Therefore, a higher  $SEN$ ,  $SPE$ , and  $ACC$  means that a heart disease classification system has a better performance.

TABLE II. CONFUSION MATRIX

Ground Truth	Prediction		
		Positives	Negatives
	Positives	True positives TP	False negatives FN
	Negatives	False positives FP	True negatives TN

$$SEN = \frac{TP}{TP + FN} \quad (9)$$

$$SPE = \frac{TN}{TN + FP} \quad (10)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

In this study, experiments were designed with the difference of training and testing datasets. In particular, it has two phases of training dataset and testing dataset scenario. In the first situation of inter-patient, DS1 for training and DS2 for testing or DS2 for training and DS1 for testing are implemented in the classifier. The second situation of intra-patient is that ADB for training and DS1 or DS2 for testing were applied to the classifier. Therefore, the results of the classification performance are evaluated using the confusion matrix as described in Table III.



TABLE III. EXPERIMENTAL RESULT IN DIFFERENTIAL DATASET FOR TRAINING AND TESTING SCENARIO

Training dataset	Testing dataset	SEN	SPE	ACC
DS1	DS2	90.35%	90.18%	90.24%
DS2	DS1	88.28%	87.74%	87.50%
ADB	DS1	100%	100%	100%
ADB	DS2	100%	99.99%	99.99%

The proposed heart disease classification performance was obtained with the very high values. In particular, the *SEN*, *SPE*, and *ACC* in case of DS1 for training and DS2 for testing are 90.35%, 90.18%, and 90.24%, respectively. The *SEN*, *SPE*, and *ACC* in case of DS2 for training and DS1 for testing are 88.28%, 87.74%, and 87.50%, respectively. Moreover, the performance of the classifier is obtained with high values related to ADB for training and DS1 or DS2 for testing. The accuracy of the classifier is obtained with 100% in case of ADB for training and DS1 for testing.

#### IV. CONCLUSION

In this paper, a 12<sup>th</sup>-order median filter was utilized to smooth ECG signals collected from the MIT-BIH database. In order to process heartbeats, each smoothed ECG signal was segmented into heartbeats. Moreover, the STFT algorithm was employed to obtain the frequency information of the heartbeat considered as features. The features were used for classifying five types of heart disease in the multi-class SVM system with OAO. Heartbeat datasets (ADB) were divided into two groups (DS1 and DS2) for evaluation of the SVM classifier. In particular, experimental results indicated that *SEN*, *TNR* and *ACC* in case of DS1 for training and DS2 for testing were 90.35%, 90.18%, and 90.24%, respectively. While *SEN*, *TNR* and *ACC* in case of ADB for training and DS1 for testing were 100%. This means that the proposed algorithm is the effective and can be developed for clinical diagnosis of heart disease in actual applications.

#### ACKNOWLEDGMENT

This work belongs to the project grant No: T2020-03NCS. funded by Ho Chi Minh City University of Technology and Education, Vietnam.

#### REFERENCES

- [1] E. Fotiadou, J. O. E. H. V. Laar, S. G. Oei, and R. Vullings, "Enhancement of low-quality fetal electrocardiogram based on time-sequenced adaptive filtering," *Medical & Biological Engineering & Computing*, vol. 56, no. 12, pp. 2313-2323, 2018.
- [2] S.-H. Liu, C.-H. Hsieh, W. Chen, and T.-H. Tan, "ECG Noise Cancellation Based on Grey Spectral Noise Estimation," *Sensors*, vol. 19, no. 4, pp. 1-14, 2019.
- [3] S. Asgari and A. Mehrnia, "A novel low-complexity digital filter design for wearable ECG devices," *PloS one*, vol. 12, no. 4, pp. 1-19, 2017.
- [4] K. S. Kumar, B. Yazdanpanah, and P. R. Kumar, "Removal of noise from electrocardiogram using digital FIR and IIR filters with various methods," in *2015 International Conference on*

*Communications and Signal Processing (ICCSP)*, pp. 0157-0162, 2015.

- [5] T.-N. Nguyen, T.-H. Nguyen, and V.-T. Ngo, "Artifact elimination in ECG signal using wavelet transform," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 936-944, 2020.
- [6] Y. Luo, R. H. Hargraves, A. Belle, O. Bai, X. Qi, K. R. Ward, *et al.*, "A Hierarchical Method for Removal of Baseline Drift from Biomedical Signals: Application in ECG Analysis," *The Scientific World Journal*, vol. 2013, pp. 1-10, 2013.
- [7] M. K. Aneja and B. Singh, "Powerline Interference Reduction in ECG Using Combination of MA Method and IIR Notch," *International Journal of Recent Trends in Engineering*, vol. 2, no. 6, pp. 125-129, 2009.
- [8] Y. Xu, M. Luo, T. Li, and G. Song, "ECG Signal De-noising and Baseline Wander Correction Based on CEEMDAN and Wavelet Threshold," *Sensors (Basel, Switzerland)*, vol. 17, no. 12, p. 2754-2769, 2017.
- [9] R. Rodríguez, A. Mexicano, J. Bila, S. Cervantes, and R. Ponce, "Feature Extraction of Electrocardiogram Signals by Applying Adaptive Threshold and Principal Component Analysis," *Journal of Applied Research and Technology*, vol. 13, no. 2, pp. 261-269, 2015.
- [10] L. N. Sharma, S. Dandapat, and A. Mahanta, "Multichannel ECG Data Compression Based on Multiscale Principal Component Analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 730-736, 2012.
- [11] C.-C. Lin and C.-M. Yang, "Heartbeat Classification Using Normalized RR Intervals and Morphological Features," *Mathematical Problems in Engineering*, vol. 2014, p. 1-12, 2014.
- [12] M. Kropf, D. Hayn, and G. Schreier, "ECG classification based on time and frequency domain features using random forests," in *2017 Computing in Cardiology (CinC)*, Rennes, France, pp. 1-4, 2017.
- [13] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 437-448, 2013.
- [14] J. Huang, B. Chen, B. Yao, and W. He, "ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network," *IEEE Access*, vol. 7, pp. 92871-92880, 2019.
- [15] B. S. Shaik, G. V. S. S. K. R. Naganjaneyulu, T. Chandrashekar, and A. V. Narasimhadhan, "A Method for QRS Delineation Based on STFT Using Adaptive Threshold," *Procedia Computer Science*, vol. 54, pp. 646-653, 2015.
- [16] V. Gupta and M. Mittal, "QRS Complex Detection Using STFT, Chaos Analysis, and PCA in Standard and Real-Time ECG Databases," *Journal of The Institution of Engineers (India): Series B*, vol. 100, no. 5, pp. 489-497, 2019.
- [17] T.-N. Nguyen, T.-H. Nguyen, M.-H. Nguyen, and S. Livatino, "Wavelet-Based Kernel Construction for Heart Disease Classification," *Advances in Electrical and Electronic Engineering*, vol. 17, no. 3, pp. 306-319, 2019.
- [18] S. Jadhav, S. L. Nalbalwar, and A. Ghatol, "Artificial Neural Network Models based Cardiac Arrhythmia Disease Diagnosis from ECG Signal Data," *International Journal of Computer Applications*, vol. 44, no. 15, pp. 8-13, 2012.

# Computational Intelligence Towards Trusted Cloudlet Based Fog Computing

Thinh Vinh Le  
IT Department

HCMC University of Technology and Education  
Ho Chi Minh city, Vietnam  
thinhlv@hcmute.edu.vn

Tran Thien Huan

Faculty of Applied Sciences  
HCMC University of Technology and Education  
Ho Chi Minh city, Vietnam  
huannt@hcmute.edu.vn

**Abstract**—The current trend of IoT user is toward the use of services and data externally due to voluminous processing, which demands resourceful machines. Instead of relying on the cloud of poor connectivity or a limited bandwidth, the IoT user prefers to use a cloudlet-based fog computing. However, the choice of cloudlet is solely dependent on its trust and reliability. In practice, even though a cloudlet possesses a required trusted platform module (TPM), we argue that the presence of a TPM is not enough to make the cloudlet trustworthy as the TPM supports only the primitive security of the bootstrap. Besides uncertainty in security, other uncertain conditions of the network (e.g. network bandwidth, latency and expectation time to complete a service request for cloud-based services) may also prevail for the cloudlets. Therefore, in order to evaluate the trust value of multiple cloudlets under uncertainty, this paper broadly proposes the empirical process for evaluation of trust. This will be followed by a measure of trust-based reputation of cloudlets through computational intelligence such as fuzzy logic and ant colony optimization (ACO). In the process, fuzzy logic-based inference and membership evaluation of trust are presented. In addition, ACO and its pheromone communication across different colonies are being modeled with multiple cloudlets. Finally, a measure of affinity or popular trust and reputation of the cloudlets is also proposed. Together with the context of application under multiple cloudlets, the computationally intelligent approaches have been investigated in terms of performance. Hence the contribution is subjected towards building a trusted cloudlet-based fog platform.

**Keywords**— Fog Computing, Reputation, Trust, Fuzzy Logic, Cloudlet, Bio-inspired Intelligent

## I. INTRODUCTION

Recently, the IoT devices have been evolved tremendously to human life from the personal use to the enterprise purposes. Since the first introduction by Cisco in 2012, Fog computing, which consists largely of conventional networking elements such as proxy servers, routers, base station transfers etc., has been known as the effective solution to enhance the performance of the IoT application in cloud environment by placing closer to the IoT devices as well as extending the benefits of Cloud computing. In Mobile Cloud Computing (MCC) context, the term cloudlet is considered as the light-weight server to provide the specific services to the proximate users via Wi-Fi technology. The cloudlet based architecture is one of the mobile edge networks (MEN) and it can act as a fog layer or fog node in the fog computing [1], [2]. Generally, it is a three-tiered (Mobile-Cloudlet-Cloud) platform which is placed to serve the proximate users and operates in the same way as a cloud but their performance is better [3], [4]. The figure 1 presents the overall environment of fog computing including cloudlet. Because of serving as a part of cloud computing environment,

the fog computing has also shared the common security concerns such as user authentication, privacy, secured data exchange and DoS attack [5]. Due to the fact that many application domains with high social and business impact such as personal healthcare, home automation, mobile payment may use IoT or mobile devices which rely solely on trust environments. In a dynamic environment, trust has become the major concern of the service providers and the customers. Trust is also an important facilitator for successful business relationships and an important technology adoption determinant [6]. Trust can be supported by hardware, software or even as a service. For example, Trust as a Service (TaaS) should provide a single point for configuring and managing the security of cloud services from multiple providers [7]. As trust is a critical factor, trust evaluation in service also plays an important role in the cloud computing. Hence, this paper focuses on how to estimate the trust and reputation in the cloudlet-based fog computing context by using fuzzy logic and ant colony optimization.

The paper is organized as follows: the related previous works are described in section 2. Then, section 3 explains the approach to estimate the trust of cloudlet with different levels of trust before presenting the role of fuzzy logic and ACO in section 4. In section 5, the experimental validation is discussed followed by the dataset and comparison with other approaches. Finally, section 6 draws the conclusion of this paper.

## II. RELATED WORK

Trust and reputation for cloud computing have recently received an attention of the research community [8], [9]. For example, to detect dishonest feedback or unfair ratings, H. Zhang et al. [10] presented a framework based on QoS similarity of the factual and advertised values to assure the impartiality and objectivity of a Web service reputation evaluation by using measurements tool. This is established in a client site to provide an automatic approach on measuring and storing QoS values of the service. In [11], Zhang et al proposed a trust model based on domain partition for decreasing the overhead of trust management i.g. trust storage and computation. The trust values have been stored in cross-domain sliding-windows. In this research, the proposed algorithm and filter procedure are also presented to remove malicious trust evaluations as well as malicious nodes from a specific domain. The authors also used a fuzzy logic to divide the trust into five levels. Other authors also used fuzzy logic apply to rule based trust evaluation in which users asked to follow some rules and the fuzzy logic is responsible to calculate trust scores based on these rules [12]. In short, this approach presented rule-based framework for evaluating trust

based on permissions which are delivered to particular users. According to the trust score, users with their permission is allowed to access specific data file. In order to against bad mouthing attacks from malicious feedback provider in edge computing, the authors in [13] proposed a multi-source feedback based trust calculation scheme. This work is relied on the idea that global trust degree of devices consists of three parts: first is a direct trust which is based on direct interaction records among interacted devices. This part is also a subjective evaluation for QoS which is provided by edge devices; second is a feedback trust from others edge devices and third is a feedback trust of service brokers. The authors claim that this approach hybrid and reliable because it has two feedbacks factors which is integrated into edge devices evaluation. Similarly, to avoid the Collusion and Sybil attacks caused by malicious users to rate malevolent feedbacks, Aarthy et al. [14] proposed mechanism to identify users and their feedbacks based on a Trusted Third Party (TTP), which is responsible for tracking the performance of cloud service providers, for producing a trustworthy and reliable feedback service for cloud storage. The TTP also has been used in [15] to prevent sensitive information leaking from the cloud service provider and data owner in the context of big data system. The authors introduced three-level definition of the verification, threat model, corresponding trusted policies based on different roles for outsourced big data system in cloud.

### III. PROPOSED TRUST BASED REPUTATION

In the cloud environment, the end user may suffer from bad services which are provided by the dishonest service providers. In fact, the quality of service in the cloud is not always right as the cloud providers advertised. The user pays for what they use but they are not aware of just how the quality of service is. Hence, the remaining challenge in the cloud is to help the user to specify the high quality of services. Along with the availability and reliability, the security and performance also are the important factor for assessing the quality of cloud service providers. Somehow, the requirement of users may force the providers to optimally balance these factors [16]. Take healthcare context for an example, the privacy and security are primary elements because of saving sensitive data. On the other hand, the normal users like student just need the service with high performance for their requesting or processing data. In order to assist the user to select the worthy service, the trust value is one of scales to estimate and identify the suitable services. In addition, the reputation is known as a subjective assessment of a cloud service for presenting public consensus towards a subject's attitude, expertise and reliability of a specific context [17]. Therefore, we have used trust-based reputation as a mechanism to assess the trustworthiness of the object in the cloud context. In this paper, we extend the work in [18] and assume that cloudlets are mentioned as fog nodes and a cloudlet network includes a cloudlet parent to manage its children as presented in Figure 1.

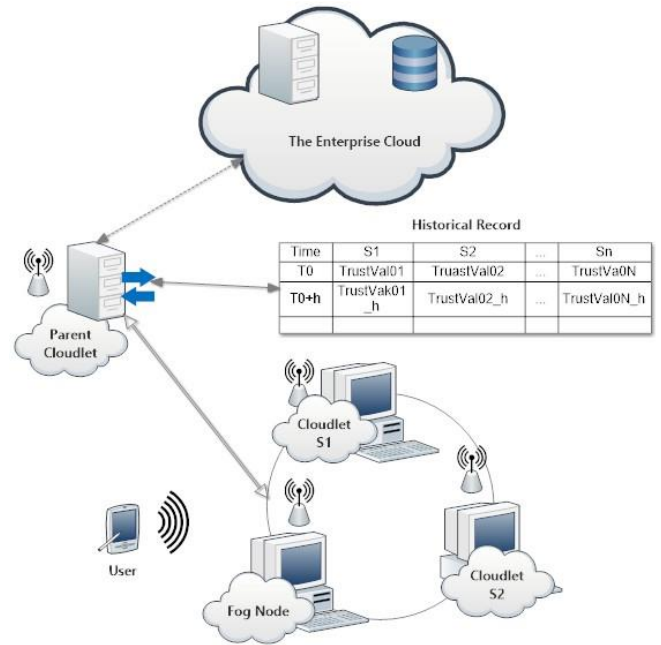


Figure 1. Cloudlet Network

In this model, for calculating the trust-based reputation value, the parent is responsible for tracking and gathering data, namely feedback, security, performance information of its children. Because the quality and ability of a particular cloudlet service provider is reflected by its trust and reputation, hence the parent cloudlet collects the reputation-based feedback as well as checks the quality of service of others in the cloudlet network. The aim of this approach is to enable the user to find a suitable service based on the result which is returned by the parent cloudlet. The parent also uses a historical record, which is saved the lasted values and added the newest value after the current session is terminated, to store the service information. The parent can use this record to keep track of the change of trust value for delivering a recommended suitable cloudlet service to the user. The reputation of each cloudlet repeatedly changes due to the genesis of a new service or replacing one old service. Moreover, the trusted third party (TTP) can play as a role of this kind of trusted cloudlet in the public cloud environment. Because of providing the primitive security level, trusted platform module (TPM) is also assumed to be mounted to all involved parties.

As presented in Figure 2, the user firstly sends his request to the cloudlet network to ask for his specific services at time  $t$ . In the same time  $t$ , the parent processes the request by searching the trust value of other cloudlets. After re-working out the trust value, the suitable cloudlet service providers are recommended to the user. In user side, after using the service, he leaves the feedback for calculating the reputation of cloudlet. However, the reputation based feedback has its own drawback as listed in [19], [5]. Admittedly, the private service provider would be dishonest of their service quality or pay to bad customers for giving high scores to their services or leave low scores to their competitors in order to gain more accessing or advertisement charges. Consequently, the end user may confuse to select fitting services for their purposes.

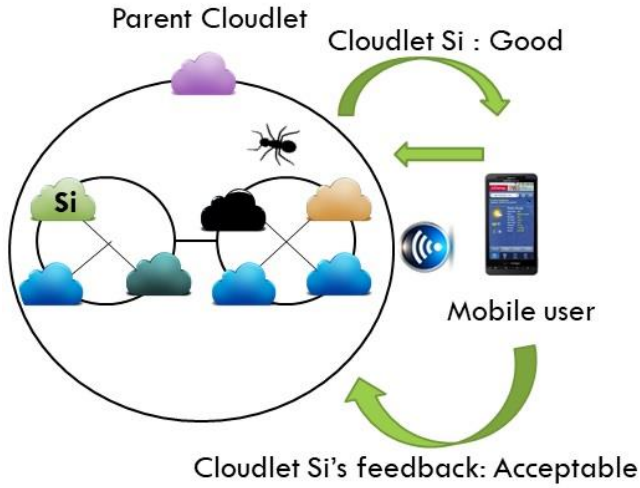


Figure 2. Fuzzy Logic and ACO for evaluating trust value

In this part, we present a model to estimate the trust value of cloudlet network based on either the reputation-based quality and uncertainty or the reputation-based feedback as presented in equation 1. In this proposed model, depending on the user's requirement, the parent cloudlet work as a filter to recommend the right service to the user.

$$\begin{aligned} TrustValue = & ReputationbasedFeedback \\ & + ReputationbasedQuality \\ & + ValueUncertainty \end{aligned} \quad (1)$$

Depending on the application type [4], the quality of service may varies. Thus, in the context of cloudlet, we focus on two prime factors of QoS at runtime such as Performance (Per) and Availability (Av). In term of security, three important pillars of security, namely Confidentiality, Integrity and Authentication (C.I.A) also involved. In addition, the uncertainty value has been simulated by using the most used pseudorandom number operators (i.e linear congruential generators) as presented in [20], [21]. By supporting machine authentication, data protection and remote attestation [22], we assume the existence of TPM as the important parameter to calculate the reputation based quality. In cloudlet network we use  $n$  cloudlets  $S_i \in (S_1 \dots S_N)$  and  $(R_{S1} \dots R_{SS})$  is the reputation-based user feedback of  $S_i$  in  $t$ . The equation 2 shows that  $R_{Si}$  is calculated by the correct feedback (cf) and the incorrect feedback (icf).

$$R_{Si} = \delta(cf) + (1 - \delta) \quad (2)$$

Where  $\delta$  is a reputation weight factor and  $0 < \delta < 1$ . We use  $R_{QoS1} \in (R_{QoS1} \dots R_{QoSS})$  as the reputations-based quality of the cloudlet  $S$  at time  $t$  and  $R_{QoS}$ , in equation 3, is composed of three parameters (Availability (Av), Performance (Per), and Secure Parameters (SP))

$$R_{QoS} = \omega_1(Av) + \omega_2(Per) + \omega_3(SP) \quad (3)$$

Where  $\omega_1$ ,  $\omega_2$  and  $\omega_3$  are the weight factors and  $\omega_1 + \omega_2 + \omega_3 = 1$ .

Availability (Av) =  $A_{req}/N_{req}$ ;  $A_{req}$  refers to the number of accepted requests and  $N_{req}$  is the total number of submitted requests. Performance (Per) =  $ts*bs/\max(bs)$  is the performance of a cloudlet in term of time efficiency, which is calculated by the bandwidth and the responded time of each

cloudlet; denoted by  $b_s$  and  $t_s$  respectively. Security Parameters include the Confidentiality (C), Integrity (I), Authentication (A) and TPM's existence (T). We assume that each child cloudlet, which works as a service provider, must comply with all the security features, like  $C \cap I \cap A \cap T$ . Therefore, C, I, A and T are considered as a single parameter (SP). Finally, the trust value (TrustVal) of a particular child cloudlet ( $S_i$ ) at time  $t$  is calculated by  $R_{Si}$  and  $R_{QoS}$  in (2) and (3) respectively combined with the uncertainty  $R_U$  Where  $\alpha$ ,  $\beta$ ,  $\gamma$  are the weight factors and  $\alpha + \beta + \gamma = 1$  as presented in equation (4)

$$TrustVal = \alpha(R_{QoS}) + \beta(R_S) + \gamma(R_U) \quad (4)$$

These values are updated to the historical record and the child cloudlet with its proper value is recommended to the user. In table 1, we use fuzzy logic to present a possible degree structure of all cloudlet nodes.

TABLE I. DEGREE STRUCTURE OF CERTAINTY ELEMENTS

Level	Scale	Status	Av & Per	$R_{Si}$	SP	Trust
1	<1	Very Poor (VP)	√	√	√	√
2	0.75-1.75	Poor(Pr)	√	√	√	√
3	1.5-3.5	Acceptable(Ac)	√	√	√	√
4	3.25-4.5	Satisfactory(Sa)	√	√	√	√
5	>4.5	Highly Satisfactory(Hs)	√	√	√	√

#### IV. BI-FOCAL: FUZZY LOGIC AND ANT COLONY OPTIMIZATION

In this paper, Fuzzy Logic System (FLS) and bio-inspired intelligence are explored to calculate the trust-based reputation for cloudlets with a reasonably high-ranking accuracy.

##### A. Fuzzification

In general, FLS, a rule based system [23], resolves difficult simulated problems with variety of input/output parameters based on the mathematical model. The process of converting the input crisp sets into fuzzy ones is the commencement of FLS. "Fuzzification" is the name to call the followings: membership functions and linguistic values. The crisp input data including Av, Per,  $R_{Si}$  and SP is converted by the fuzzier into fuzzy sets containing the membership functions (eq. 5) and linguistic variables (Vp, Pr, Ac, Sa and Hs) as presented in table 1 and figure 3.

$$\mu = \begin{cases} 1 - \frac{x-c}{r-c} & (c < x < r) \\ 1 & (c = x) \\ 1 - \frac{c-x}{c-l} & (l < x < c) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Where:  $l$ ,  $c$  and  $r$  refer to left, center and right respectively. For example, on  $x$  axis in figure 3, if we have ( $l = 1.5$ ;  $c = 2.5$ ;  $r = 3.5$ ) then  $\mu_{Ac}(x) = 0.5$ .



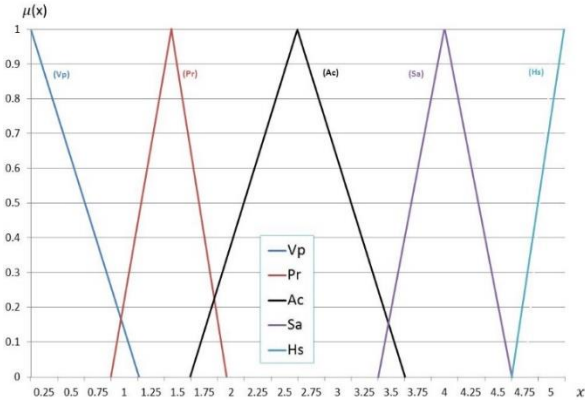


Figure 3. Membership input functions of linguistic variables

Additionally, it uses fuzzy model to extract compelling rules from quantitative or linguistic values in relational and transactional databases [24]. Subjective trials are used to boost the number of rules freely by the user. Thus, many non-compelling rules might be created resulting in the increment of computation cost. For instance, it takes more time for the application to process the algorithm as a result of complication and variety of number of rules. The target of association rule mining in fuzzy logic is to search for all rules which have the degree of support and confidence greater than the degree of support ( $\min\_sup$ ) and confidence ( $\min\_conf$ ) set by the user (i.e. administrator or super-user of cloud/cloudlet). The form of the association rule is demonstrated as "If X is A Then Y is B" whilst the support of association rule refers to the degree to which relationship exists in the database,

$$Support_{rule}(A \rightarrow B) = \frac{|\sum_{x_p \in Class(y_p)} \beta_{AB}(x_p)|}{|N|} \quad (5)$$

and the confidence of association rule is the probability that if X then Y,

$$Confidence_{rule}(A \rightarrow B) = \frac{|\sum_{x_p \in Class(y_p)} \beta_{AB}(x_p)|}{|\sum_{x_p \in Class(y_p)} \beta_A(x_p)|} \quad (6)$$

Where,

- $|N|$  is the number of transactions
- $\beta_A(x_p)$  is the compatibility degree of transaction  $x_p$  with the antecedent part A
- $\beta_{AB}(x_p)$  is the compatibility degree of transaction  $x_p$  with the antecedent and consequent of the rule  $A \rightarrow B$
- $y_p$  represents  $p^{th}$  of p consequent.

Developing the fuzzy rules known as the min-max rule for conjunctive (AND) and disjunctive (OR) reasoning is exhausting. These rules are arbitrary, which is difficult to figure out the sufficient number of rules to tackle a specific issue from all angles and to make the related robust membership functions. Besides, the fuzzy logic requires a considerable amount of data to formulate fuzzy membership functions. The decisions can also be interpreted in a number of different possibilities [25], [26] resulting in a tradeoff between performance and robustness. Since the rules and

ranges of value cannot be defined automatically, the fuzzy logic is also considered as a static method. Hence, to be more dynamic and adaptive, the paper solicits another computationally intelligent approach (e.g. ant colony optimization) to consolidate the measure of trust.

### B. Modeling trust with ant colony system: mathematical perspectives

According to the reputation, the ant colony system is modeled to distribute the load of the cloudlets. It also expresses a reverse way to identify the most trusted cloudlet that will have multiple numbers of service requests with initial and final time intervals. In this paper, the ant colony optimization is used as a trust or reputation search strategy while distributing the service load. The cloudlet nodes are represented as an undirected graph  $G(V, E)$ , where  $V$  is the set of all child cloudlets in the specific cloudlet network whose reputation has to be measured.  $E$  is the network set which connects to each cloudlet. After dividing the cloudlet network with effective groups, ants can be assigned to each group randomly. Each ant will search for trusted entropy from random assignments. There might be several explorations of the trusted path in the form of different search algorithms. The investigation of a trusted and optimal search path requires modifying the pheromone matrix in regard to loss function work as an evaporation mechanism. In this context, from the ant colony perspectives, group of different colonies can be used for different cloudlet in the same location. The behavior of ants in one colony will be influenced by the solution information received from other colonies, where a pheromone is added to the colony edges that belong to the best solutions of the group of colonies. The pheromone trails on the edges of the best solutions are updated adaptively in react to determined weights. An extra quantity of pheromone has been deposited on the edges of these solutions accordingly. The equation 8 demonstrates the trusted optimal path (Equation 8).

$$Rt_{avg} = \frac{1}{|S|} \sum_{h \in S} Rt_{opt}^h \quad (8)$$

While fulfilling the final roadmap of the proposed model, the model assumes recursive solutions:

- Overall average reputation and trust  $Rt_{avg}$ ,
- $h \in S$ , where  $S$  is the optimal search solution and  $h$  is the dynamic edge value of trust or reputation for each edge of cloudlet connection.
- $Rt_{opt}^h$  is the value of optimal trust with the dynamic edge value  $h$ .
- $\tau_0$  is the initial amount of pheromone.
- The evaporation rate,  $\rho$ , is a parameter in the range  $[0, 1]$  that regulates the lessening of pheromone on the edges.

Pheromone trail  $S_{ij}(ij)$ , where the first part is the suffix of  $s$  and the other part  $ij$  is the factor of  $S$  which means pheromone trail  $s$  from node  $i$  to  $j$  for suffix and next  $ij$  is the factor that is the actual numerical value of pheromone which is approximated deposited or evaporated ; for factor  $ij$  part the edge will be visible if it is high, which means other ants will follow it.

The equations 9 present the compound mathematical expression. In equation 9,  $\tau_0$  is finally replaced by the



optimum trust; this is in denominator to make sure that the optimal value is always a lessened value with power of time.

$$\begin{aligned}\tau_{ij}(t+1) &= (1-\rho)\tau_{ij}(t) + \rho\tau_0 \\ \tau_{ij(t+1)} &= (1-\rho)\tau_{ij}(t) + \rho\left(\frac{1}{Rt_{opt}}\right)\end{aligned}\quad (9)$$

Where,  $\tau_{ij}$  is the edge value of pheromone for graph from cloudlet  $i$  to cloudlet  $j$  and  $Rt_{opt}$  is the value of optimal trust. This expression has the influence of increasing the amount of pheromone, which represents reputation and trust, on edges associated with solutions as stated by the quality of trust and reputation and to the current convergence state of the cloudlet information exchange.

Finally, the value of affinity (refers equation 10) of trust for  $M$  number of ant colonies and  $N$  number of cloudlets of specific cloudlet network can be calculated:

$$affinity(M, N) = 1 - \frac{\sqrt{\frac{\sum_{x \in P(M, N)} \frac{\sum_{i=1}^P P(x, M) s(x, i)}{P(x, N)}}{|PI(M, N)|}}}{|PI(M, N)|} \quad (10)$$

In general scenario, the proposed approach is not only applied to a cloudlet network but also to multiple cloudlet networks. The complete participation for all cloudlets in different cloudlet networks can indicate  $PI(M, N)$ ,  $P(x, M)$  can indicate individual information exchange towards trust gain.  $P(M, N)$  is the total information exchange for  $M$  number of ant colonies and  $N$  numbers of cloudlets;  $s(x, i)$  is the optimal search function from the initial position of dynamic edge  $h$  and  $x$  is the part of edge which will be searched through search function. This affinity is a probabilistic numerical value, which indicates trust gain from a base cloudlet/cloudlet network to a destination token. The choice of the affinity value is an inductive approach to demonstrate the gain of trust value for a specific location-based cloudlet. If more locations and cloudlets are connected, the affinity value also becomes recursive.

In general, the proposed approach can be elaborated as:

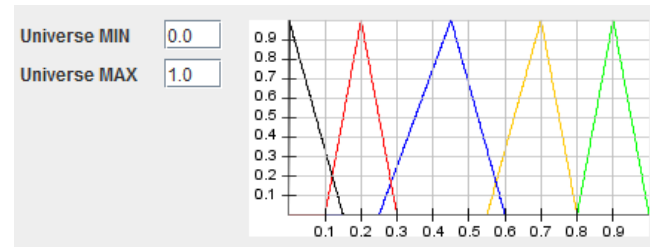
- Step 1: Allocate the ants randomly in the solution space. Each ant in the solution space correlates with the possible combination of the different multiple cloudlet parameters.
- Step 2: Asses the value of the loss function
- Step 3: From the Fuzzy rule base: The rule base has been formed by utilizing the past experiences. (table 1)
- Step 4: Calculation of the resultant location utilizing the fuzzy rule base, evaluate to the next best location that the ant can occupy.
- Step 5: Update the ant pheromone for cloudlet nodes
- Step 6: Estimate the movement of the ants
- Step 7: Check if the termination criterion is reached or not. If reached, then stop the algorithm or otherwise move to step 2.

## V. EXPERIMENTAL VALIDATION AND RESULT

Thanks to Wi-Fi hotspot servers, the cloudlet-based fog computing is considered as a resource-rich, well-connected and powerful computer installed in the middle of public infrastructure with the connectivity to a cloud server. Therefore, cloudlet-based fog architecture also succeeds the

Quality of Service, Security features of the Cloud. In this part, the performance of various relevant trust and reputation models is discussed in terms of accuracy, path length and energy consumption. We use TRMSim-WSN [27], which is a Java-based trust reputation models simulator, to make a comparison among discussed models. The main purpose of TRMSim-WSN is to provide an easy method to test a trust and reputation model over WSNs and enable users to compare their own model against others. In this part, we consider comparing the proposed approach to LFTM, which includes BRTM-WSN, and PeerTrust in dynamic WSNs environment. We keep the network configuration similar to these models and also adjust the parameters to match with the proposed approach settings as presented in figure 4. In the proposed model, for example, we do not rely mainly on the user's feedback to modify/update the pheromone value and the client is supposed to receive the suitable requested service from the cloudlet master. Hence, we set the threshold of punishment and transition is equal to maximum and minimum, respectively. Regarding to fuzzy logic, we also change the model of membership function to triangular instead of trapezoidal of LFTM, to match with figure 3. The data of the following graphs is collected by the first 10 values when the simulator is started.

a) Network configuration



b) Membership function settings

Figure 4. TRMSim-WSN configuration settings.

### A. Accuracy

The term accuracy refers to the percentage of number of times for successful selecting trustworthy nodes. In graph A, it shows that PeerTrust model has the most stable accuracy. Its accuracy oscillates in the smallest range ( $\Delta_{Peertrust} = maximum - minimum = 97.83\% - 74.76\% = 23.07\%$ ). Although LFTM has the highest accuracy point (98.57%) but

its  $\Delta_{LFTM} = 37,75\%$  is higher than others. Similar to PeerTrust, the proposed model also has a stable accuracy and obtain an acceptable accuracy with a maximum and minimum value is 98.11% and 72.42% (Figure 5).

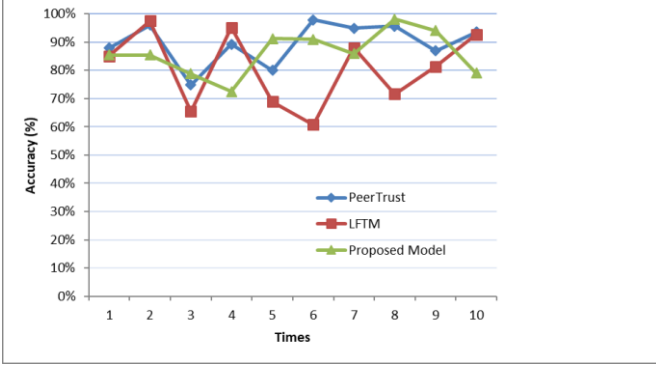


Figure 5. Accuracy comparison

### B. Path Length

According to [28], [29], path length is considered as the average hops leading to the trustworthy nodes which are selected by the users in the network. It assumes that the smaller number, the better performance. As shown in graph B, the proposed model has the best performance when it has the lowest path length point (2.09) in comparison with the worst performance of PeerTrust (7.26). The result of LFTM and proposed model in this comparison prove that the bio-inspired technique is an effective solution for evaluating the trust and reputation in computing network (Figure 6).

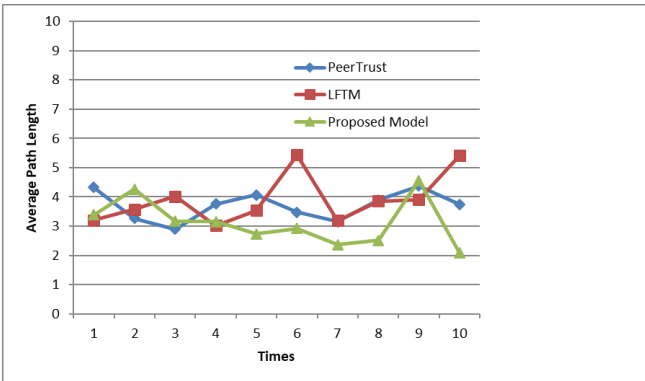


Figure 6. Path Length comparison

### C. Energy consumption

Energy consumption is the overall energy consumed by all involved entities, such as the clients and remote nodes. The figure 7 shows the truth that there is the trade-off between security and performance; and the more security the model supports the more energy they consume. For example, LFTM is considered as the most effective model but it consumes the highest energy. Moreover, the graph also demonstrates that the proposed model is suitable for the mobile user in mobile cloud computing environment where the mobile device is known as a resource-constraint device.

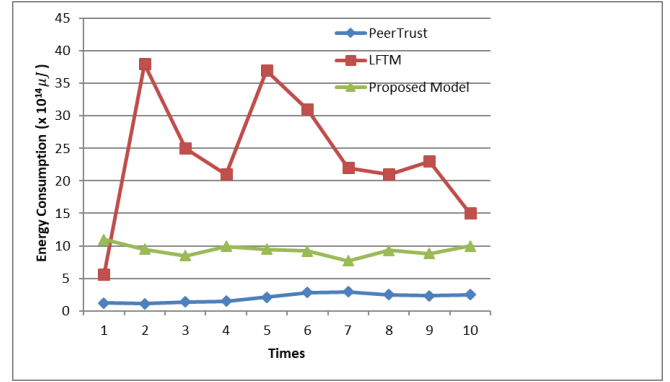


Figure 7. Energy consumption comparison.

In short, this part shows the overview of the difference between the proposed approach and the most known models. Although each model has its own pros and cons, they are considered as the effective model to assess the trust and reputation. According to the comparison data, it is worth noting that the proposed approach has obtained the acceptable results in terms of accuracy, shortest path and energy consumption. In the following section we draw the conclusion of this paper.

## VI. CONCLUSION

In the cloud environment, classifying the suitable services for the purpose is significant. While the intrusion detection and associated measures cannot guarantee trust and validation of multiple intelligent IoT devices under the cloudlet. The main purpose of this paper is to evaluate the trust value of multiple cloudlets under uncertainty by proposing specific hybridization of computationally intelligent measures of trust both from the perspectives of the users and the quality of service. The trust evaluation has used fuzzy components which are based on linguistics decisions of human users. Parallel, ant colony can inspect the performance and quality aspects of messages in cloudlet transmission. For the future work, to handle the uncertainty value more precisely, we are going to use the Monte Carlo simulation.

## REFERENCES

- [1] C. Dsouza, G.-J. Ahn, and M. Taguinod, "Policy-driven security management for fog computing: Preliminary framework and a case study," in *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 2014, pp. 16–23.
- [2] V. Cardellini, V. Grassi, F. L. Presti, and M. Nardelli, "On QoS-aware scheduling of data stream applications over fog computing infrastructures," in *2015 IEEE Symposium on Computers and Communication (ISCC)*, Larnaca, Jul. 2015, pp. 271–276, doi: 10.1109/ISCC.2015.7405527.
- [3] S. Simanta, K. Ha, G. Lewis, E. Morris, and M. Satyanarayanan, "A reference architecture for mobile code offload in hostile environments," in *International Conference on Mobile Computing, Applications, and Services*, 2012, pp. 274–293, Accessed: Jun. 20, 2016. [Online]. Available: [http://link.springer.com/chapter/10.1007/978-3-642-36632-1\\_16](http://link.springer.com/chapter/10.1007/978-3-642-36632-1_16).
- [4] L. A. Tawalbeh, F. Ababneh, Y. Jararweh, and F. AlDosari, "Trust delegation-based secure mobile cloud computing framework," *Int. J. Inf. Comput. Secur.*, vol. 9, no. 1/2, p. 36, 2017, doi: 10.1504/IJICS.2017.10003598.
- [5] R. Mahmud, R. Kotagiri, and R. Buyya, "Fog Computing: A Taxonomy, Survey and Future Directions," in *Internet of Everything*, B. Di Martino, K.-C. Li, L. T. Yang, and A. Esposito, Eds. Singapore: Springer Singapore, 2018, pp. 103–130.

- [6] J. Lansing and A. Sunyaev, "Trust in Cloud Computing: Conceptual Typology and Trust-Building Antecedents," *SIGMIS Database*, vol. 47, no. 2, pp. 58–96, Jun. 2016, doi: 10.1145/2963175.2963179.
- [7] J. Huang and D. M. Nicol, "Trust mechanisms for cloud computing," *J. Cloud Comput. Adv. Syst. Appl.*, vol. 2, no. 1, p. 9, Apr. 2013, doi: 10.1186/2192-113X-2-9.
- [8] L. Zhou, F. Zhang, and G. Wang, "Using asynchronous collaborative attestation to build a trusted computing environment for mobile applications," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and SmartCityInnovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, Aug. 2017, pp. 1–6, doi: 10.1109/UIC-ATC.2017.8397459.
- [9] S. B. Hosseini, A. Shojaei, and N. Agheli, "A new method for evaluating cloud computing user behavior trust," in *2015 7th Conference on Information and Knowledge Technology (IKT)*, Urmia, Iran, May 2015, pp. 1–6, doi: 10.1109/IKT.2015.7288735.
- [10] H. Zhang, Z. Shao, H. Zheng, and J. Zhai, "Web Service Reputation Evaluation Based on QoS Measurement," *Sci. World J.*, vol. 2014, p. e373902, Apr. 2014, doi: 10.1155/2014/373902.
- [11] P. Zhang, Y. Kong, and M. Zhou, "A Domain Partition-Based Trust Model for Unreliable Clouds," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2167–2178, Sep. 2018, doi: 10.1109/TIFS.2018.2812166.
- [12] P. Rathi, H. Ahuja, and K. Pandey, "Rule based trust evaluation using fuzzy logic in cloud computing," in *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, Sep. 2017, pp. 510–514, doi: 10.1109/ICRITO.2017.8342481.
- [13] J. Yuan and X. Li, "A multi-source feedback based trust calculation mechanism for edge computing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Honolulu, HI, Apr. 2018, pp. 819–824, doi: 10.1109/INFOCOMW.2018.8406900.
- [14] D. K. Aarthy, M. Aarthy, K. A. Farhath, S. Lakshana, and V. Lavanya, "Reputation-based trust management in cloud using a trusted third party," in *2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM)*, Chennai, Mar. 2017, pp. 220–225, doi: 10.1109/ICONSTEM.2017.8261418.
- [15] J. Zhan, X. Fan, L. Cai, Y. Gao, and J. Zhuang, "TPTVer: A trusted third party based trusted verifier for multi-layered outsourced big data system in cloud environment," *China Commun.*, vol. 15, no. 2, pp. 122–137, Feb. 2018, doi: 10.1109/CC.2018.8300277.
- [16] T. L. Vinh, H. Cagnon, S. Bouzeffrane, and S. Banerjee, "Property-based token attestation in mobile computing: Property-based token attestation in mobile computing," *Concurr. Comput. Pract. Exp.*, p. e4350, Oct. 2017, doi: 10.1002/cpe.4350.
- [17] Y. Zuo, "Reputation-based service migration for moving target defense," May 2016, pp. 0239–0245, doi: 10.1109/EIT.2016.7535247.
- [18] T. Le Vinh, "Security and Trust in Mobile Cloud Computing," PhD Thesis, Conservatoire national des arts et metiers-CNAM, 2017.
- [19] T. H. Noor, Q. Z. Sheng, and A. Alfazi, "Reputation Attacks Detection for Effective Trust Assessment among Cloud Services," Jul. 2013, pp. 469–476, doi: 10.1109/TrustCom.2013.59.
- [20] K. Entacher, A. Uhl, and S. Wegenkittl, "Linear Congruential Generators for Parallel Monte Carlo: the Leap-Frog Case.," *Monte Carlo Methods Appl.*, vol. 4, no. 1, pp. 1–16, 1998, doi: 10.1515/mcma.1998.4.1.1.
- [21] J. E. Gentle, *Random Number Generation and Monte Carlo Methods*. New York: Springer-Verlag, 2003.
- [22] W. Arthur and D. Challener, *A Practical Guide to TPM 2.0: Using the Trusted Platform Module in the New Age of Security*. Apress, 2015.
- [23] L.-X. Wang, *A Course in Fuzzy Systems and Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1997.
- [24] M. S. Khan, M. Mueyba, C. Tjortjis, and F. Coenen, "An effective fuzzy healthy association rule mining algorithm (FHARM)," *databases*, vol. 4, no. 5, p. 14, 2007.
- [25] K. R. Sasikala and M. Petrou, "Generalised fuzzy aggregation in estimating the risk of desertification of a burned forest," *Fuzzy Sets Syst.*, vol. 118, no. 1, pp. 121–137, Feb. 2001, doi: 10.1016/S0165-0114(99)00064-0.
- [26] M. Shamim, S. Enam, U. Qidwai, and S. Godil, "Fuzzy logic: A 'simple' solution for complexities in neurosciences?," *Surg. Neurol. Int.*, vol. 2, no. 1, p. 24, 2011, doi: 10.4103/2152-7806.77177.
- [27] F. G. Mármol and G. M. Pérez, "TRMSim-WSN, trust and reputation models simulator for wireless sensor networks," in *Communications, 2009. ICC'09. IEEE International Conference on*, 2009, pp. 1–5, Accessed: Jul. 18, 2017. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/5199545/>.
- [28] F. Gómez Mármol and G. Martínez Pérez, "Providing trust in wireless sensor networks using a bio-inspired technique," *Telecommun. Syst.*, vol. 46, no. 2, pp. 163–180, Feb. 2011, doi: 10.1007/s11235-010-9281-7.
- [29] H. Marzi and M. Li, "An Enhanced Bio-inspired Trust and Reputation Model for Wireless Sensor Network," *Procedia Comput. Sci.*, vol. 19, pp. 1159–1166, 2013, doi: 10.1016/j.procs.2013.06.165.

# Traditional Method Meets Deep Learning in an Adaptive Lane and Obstacle Detection System

Van-Tin Luu

*Faculty of High Quality Training  
HCMC University of Technology  
Education*

Ho Chi Minh City, Viet Nam  
16119048@student.hcmute.edu.vn

Viet-Cuong Huynh

*Faculty of High Quality Training  
HCMC University of Technology  
Education*

Ho Chi Minh City, Viet Nam  
16119003@student.hcmute.edu.vn

Vu-Hoang Tran

*Faculty of Electrical and Electronics  
Engineering  
HCMC University of Technology  
Education*

Ho Chi Minh City, Viet Nam  
hoangtv@hcmute.edu.vn

Trung-Hieu Nguyen

*Faculty of Vehicle and Energy Engineering  
HCMC University of Technology Education*

Ho Chi Minh City, Viet Nam  
hieuntr@hcmute.edu.vn

Thi-Ngoc-Hieu Phu

*Faculty of Electrical and Electronics Engineering  
HCMC University of Technology Education*

Ho Chi Minh City, Viet Nam  
hieuptn@hcmute.edu.vn

**Abstract**—Recently, a lot of researches and applications regarding to autonomous vehicle have been invested and developed. These applications not only reduce the risk of traffic accidents but also bring convenience to the driver. In these applications, lane and obstacle detection are indispensable tasks. Therefore, in this paper, an adaptive lane and obstacle system has been proposed to solve the above issues. In particular, our model not only detects the lanes but also avoids the obstacle on the road. Our system is tested on the simulated environment using Unity software then embedded on a 1:7 RC vehicle and tested on a small driving environment. The experimental results demonstrate the effectiveness and robustness of the proposed model in many challenging situations.

**Keywords**—Lane and Obstacle detection, deep learning, image processing, light-weight model.

## I. INTRODUCTION

Many methods were proposed [1] [2] [3] [4] [5] to develop the lane detection systems. They can be divided into two categories: (1) image processing and (2) deep learning methods. With image processing methods, although the achieved speed and response time are very fast, they are only suitable in some simple and specific environments. Meanwhile using deep learning methods can resolve the problems in more complicated environment, but their speed and response time depend on the processing hardware and designed framework. Besides, deep learning methods usually require a lot of labeled data, which takes not only time but also effort to collect and process. Therefore, in this paper, we combine both image processing and deep learning methods into a unified system to take advantage of their own benefits.

Most of the existing systems [6] [7] [8] [9] focused on the simple cases such as highway driving. A few recent researches [10] [4] developed algorithms for more complex and challenging environments such as urban areas. Hence, they have to face more difficult issues. Most of them come from cluttered image background such as, curved lanes, obstacles,

and cast shadows from multiple objects on the road. So, we need a powerful method to handle those factors.

Recently, authors in [11] [12] applied deep learning-based segmentation networks to perform lane segmentation. These approaches are effective for the above issues, need less computation time and achieve the better results, but they may create the segmentation results with broken and non-continuous sections. This makes them difficult to merge lane segments belonged to the same lane as a unified group for other semantic applications. In [13], the authors showed a good solution to handle effectively most of the previous problems, but it is not real-time and cannot be used for embedded system. In this paper, we inherit some idea from the upon system, but we improve them to be suitable for our hardware and have a real-time system.

The paper is organized as follows: Section II describes the proposed method. Section III shows the experimental results and comparison. In section IV, we conclude the paper.

## II. PROPOSED METHOD

The overall process of our algorithm is illustrated in Fig. 1, it consists of two steps that run in parallel: (1) RGB processing and (2) depth processing. In RGB processing step, we would like to detect the lane from RGB image by combining both traditional image processing and deep learning methods. Based on the detected lanes from the first step, we then use the depth map, in depth processing step, to filter out the background information and detect the obstacle on the road. Our aim is to process both of lane and obstacle detection in order to identify the drivable area. Next, we detail them.

### A. RGB processing

In our system, in order to not only boost the performance but also reduce the computation time, we combine both deep learning and traditional image processing methods into a unified framework.

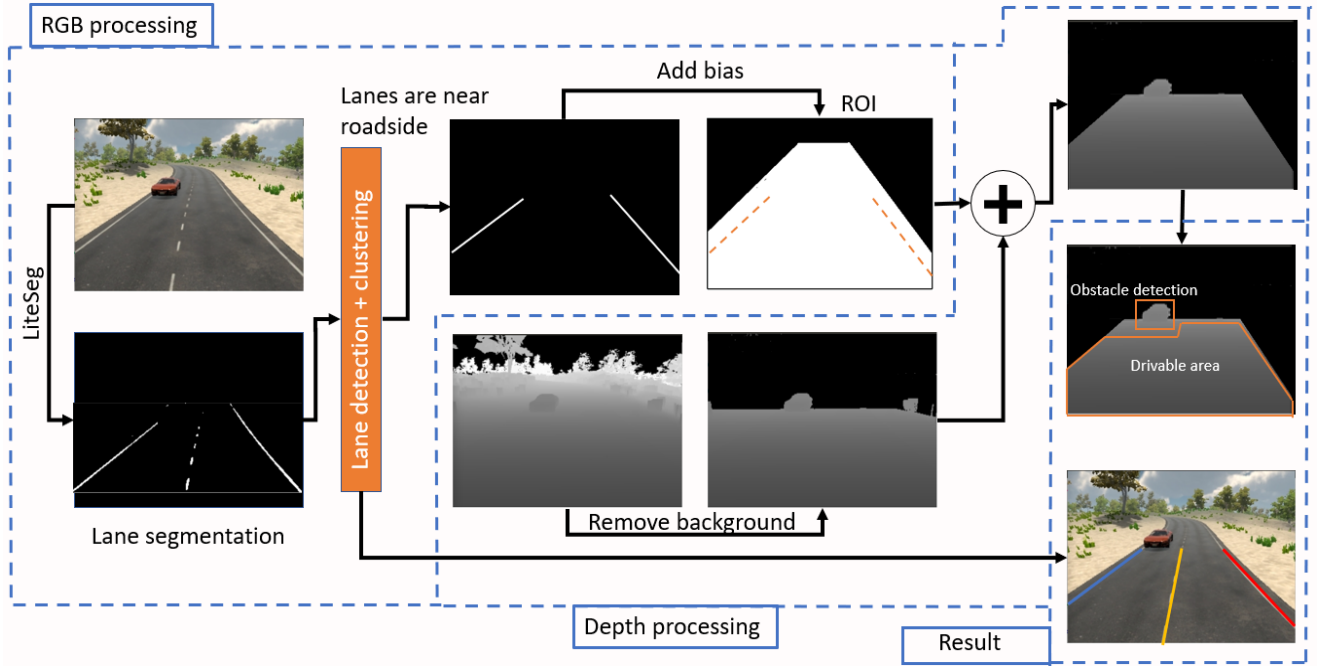


Fig 1. The proposed system.

### 1) Deep learning-based lane segmentation

We first proposed to use the LiteSeg network architecture [14], a lightweight architecture that is suitable for running in real-time situation, to extract the coarse lane segmentation. The input of network is the captured RGB image and the output is the lane segmentation map with 2 classes: lane and non-lane. Our backbone network is MobileNetV2 [15] with depth-wise and inverted residual structure. With this design, the number of weights is much smaller, the size of model is reduced, and the calculation speed is significantly increased. To solve the challenge about multi-scale information, the network employs Atrous Spatial Pyramid Pooling (ASPP) module like the method [16] with different dilation rates. And to solve the scale problem, the network is trained with multi-resolution images. The proposed architecture is given in Fig. 2.

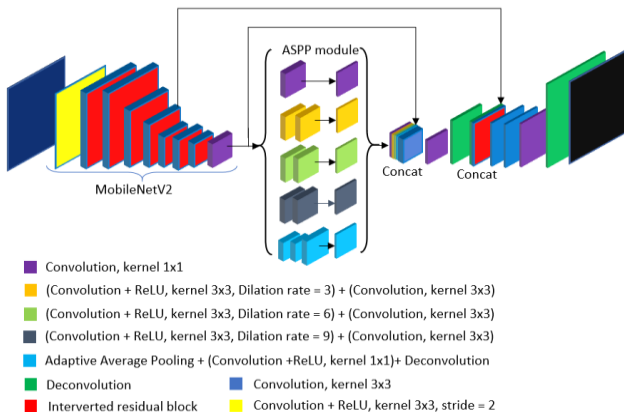


Fig 2. The details of the proposed LiteSeg.

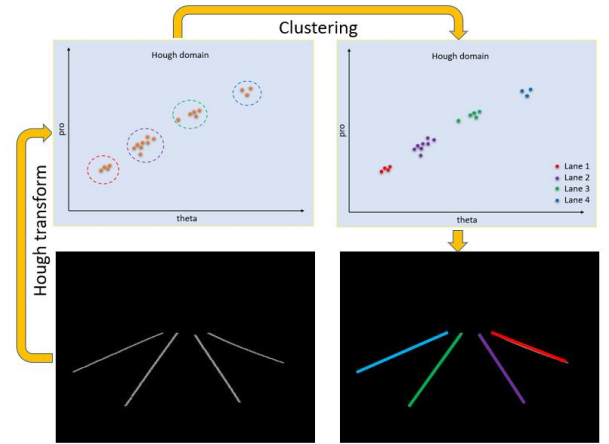


Fig 3. Hough transform based Lane detection method.

### 2) Hough transform based lane detection

However, the LiteSeg network with MobileNetV2 backbone is not strong enough to correctly deduce all lanes. There are a lot of noises and fragmentations in the obtained results, so to solve this problem we propose a Hough transform based lane detection method. By Hough transform, the Image domain is transformed to Hough domain and each lane component is signified as a Hough point as shown in Fig. 3.

Based on distances we can determine the cluster in Hough domain as shown in Fig. 4a. But if we had over a thousand points, the calculation would be very time-consuming. So, for each point ( $\rho$ ,  $\theta$ ) on the Hough domain, we extend it by drawing a circle on the binary image where ( $\rho$ ,  $\theta$ ) is the center of the circle and the radius is a fixed value  $R$  as shown in Fig. 4b. By finding the contour of the connected components using [17], the clusters are then defined automatically. After clustering, the small clusters are considered as noise and filtered out. The remaining clusters,



which is defined as lanes, will be transferred back to the image domain for curve fitting.

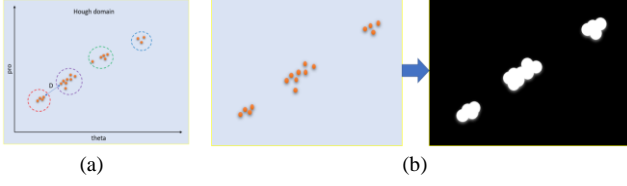


Fig 4. (a) Distances based clustering. (b) Pre-processing before clustering.

### 3) Curve fitting

To deal with the curvy lanes, we model our lane using the quadratic polynomial as shown in equation 1. The obtained candidate segments are then fitted into the lane model using Polynomial curve fitting [18]. After that, we can define the road ROI by using the obtained outermost lanes. The defined ROI will be then sent to depth processing task for further processing.

$$y = ax^2 + bx + c, \quad (1)$$

where  $(x, y)$  is pixel coordinate of lane candidates;  $a$ ,  $b$  and  $c$  are the coefficients found by Polynomial curve fitting.

### B. Depth processing

#### 1) Threshold based background subtraction

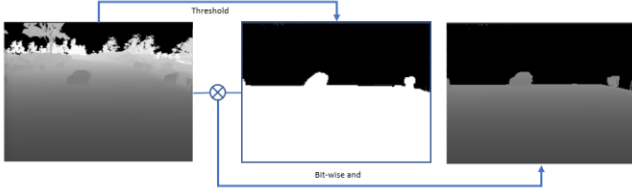


Fig 5. Background subtraction.

In order to reduce the noise from complex background, we take advantage of depth image to filter out the background. For ADAS systems, we only consider the objects at a fixed distance in front of our vehicle. For objects that are too far away, we will treat them as the background and eliminate them. As shown in Fig. 5, the depth image is first converted into 8-bit grayscale, then information that is too far from the camera is filtered using a fixed threshold. The filtered result is then combined with the road ROI, defined based on RGB image, to further filter out the information that is out of the road as shown in Fig. 6.

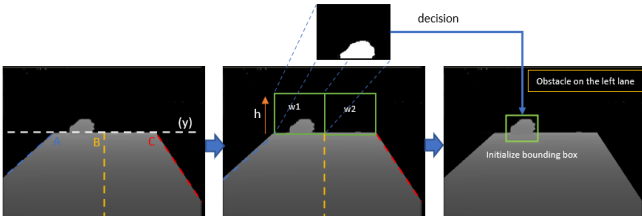


Fig 6. The details of obstacle detection.

#### 2) Obstacle detection and drivable area segmentation

Based on the obtained result of background subtraction step, we then detect the obstacle on the road to find the drivable area as shown in Fig. 6. We first divide the ROI area

into two segments, representing two lanes of the road. Then two sliding windows will slide along these segments to identify obstacles on the corresponding lanes. The obstacles will be defined if the number of non-zero pixels in these windows exceeds a pre-defined threshold. Once obstacles are detected, their bounding boxes and locations on road are extracted. By subtracting the obstacle areas out of road image, the drivable area can be then segmented.

## III. EXPERIMENTAL RESULTS

### A. Datasets:

In our system, we evaluate the performance of our system according to 2 tasks: lane detection and obstacle detection. We first choose the CamVid dataset [19], which includes 4000 training images and 826 test images, for evaluating our lane detection model. Besides, we also collect our own dataset on a small driving environment using Kinect camera embedded on 1:7 RC car. Our dataset contains 1000 labeled images which include several challenging cases for evaluating our lane and obstacle detection algorithm. The collected images are shown in Fig. 7.



Fig 7. Image acquisition.

### B. Implementation details

To speed up the training, the backbone network is initialized by the pre-trained network MobileNetV2. As aforementioned, we train our network with multi-scale images. We augment the dataset by resizing the input image into 4 different sizes: 512x1024, 768x1536, 360x640, 720x1280. We employ Adam solver [20] to optimize the softmax function with weight decay  $4 \times 10^{-5}$  and learning rate is initialized as  $10^{-7}$ . Learning rate changes after five epochs, each epoch the learning rate is calculated by Eq. 3. We train the model with 200 epoch and the batch size is 4.

$$learning\_rate = 10^{-7} \times \left( \frac{1 - epoch}{max\_epochs} \right)^{0.9}, \quad (2)$$

### C. Results

To evaluate the performance of lane detection, we use the following metrics: mean intersection over union (mean IU), mean accuracy (ACC), giga floating point operations (GFLOPs), and frames per second (FPS). Mean IU and mean accuracy are defined in [21]. We will test our system on NVIDIA Jetson TX2.

In Table I, we show the performance of our model with different sizes of input on Camvid dataset. The results show that it's better to use our model with the input size of 720x1280 or 360x640. With the larger input size, 720x1280, although the performance is a bit higher, the speed is reduced by more than 4 times compared to using input size of 360x640 as shown in Table II. Besides, in Table I, we also

compared the performance of our model with and without multi-scale (MS) training. When using MS training, the performance is boosted around 1% on Mean IU and 1% to 2% on Mean ACC.

Some results of the segmentation model on Camvid dataset, simulated environment using Unity software, and our collected dataset are given in Fig. 8, Fig. 9 and Fig. 10. The results indicate that the model adapts quite well to the new environments, even though it has not been re-trained on these environments. Also, on that dataset, we compare the proposed method with SkipNet [22], a light-weight model for lane segmentation. The results, given in Table III, show that our performance is better than SkipNet and the computing cost is also cheaper.

Some results of the proposed method, combining deep learning and traditional image processing, on our collected dataset are given in Fig. 11. As shown in the Figure, our method can still identify the curve lanes while moving with a high speed. Besides, in order to evaluate obstacle detection's performance, we use true positive rate (TPR) and false positive rate (FPR) as shown in the Table IV. The results signify that our obstacle detection method infer well at distances less than 6m.

TABLE I. SEGMENTATION PERFORMANCE.

<i>Input size</i>	<i>MS</i>	<i>Mean IU, %</i>	<i>Mean ACC, %</i>
360x640	Yes	73.71	94.1
	No	72.01	92.5
<b>720x1280</b>	<b>Yes</b>	<b>78.08</b>	<b>97.9</b>
	No	77.5	96.02
512x1024	Yes	68.44	87.1
	No	67.77	85.56
480x640	Yes	62.38	80.9
	No	59.03	76.1

TABLE II. THE NUMBER OF FLOATING-POINT OPERATIONS AND THE INFERENCE TIME.

<i>Input size</i>	<i>GFLOPS</i>	<i>FPS (FP32)</i>	<i>FPS (FP16)</i>
<b>360x640</b>	<b>2.18</b>	<b>14.5</b>	<b>20.6</b>
720x1280	8.56	3.54	5.3

TABLE III. COMPARISON BETWEEN THE PROPOSED MODEL AND SKIPNET.

<i>Method</i>	<i>Input size</i>	<i>GFLOPS</i>	<i>Mean IU, %</i>	<i>Mean ACC, %</i>
SkipNet [22]	360x640	6.2	57.02	80.05
<b>Proposed method</b>	<b>360x640</b>	<b>2.18</b>	<b>73.71</b>	<b>94.1</b>

TABLE IV. LANE DETECTION PERFORMANCE WITH DIFFERENT DISTANCES.

<i>Distance from camera to obstacle, meter</i>	<i>TPR, %</i>	<i>FPR, %</i>
<b>&lt; 6</b>	<b>99.96</b>	<b>0.009</b>
6 - 7	62.56	0.01
> 7	-	-

#### IV. CONCLUSION

This paper we proposed a unified system which can detect lanes and obstacles on the road. For detecting lane, a combination of deep learning and traditional image processing framework was proposed to reduce the data collection time and effort while maintaining the performance. The detected lanes are then used to define road and detect the

obstacles on that road by taking advantage of depth information. The visual and evaluation results show that our system work well in our testing environment with a model car. In the future, we will try to embed the system in a real car and consider more reality challenges.



Fig. 8. Lane segmentation results on Camvid dataset.



Fig 9. Lane segmentation results on the simulated environment



Fig 10. Lane segmentation results on small driving environment.



Fig. 11. Lane detection and drivable area segmentation results in small driving environment.

## REFERENCES

- [1] Jamel Baili, Mehrez Marzougui , Ameer Sboui , Samer Lahouar , Mounir Hergli , J.Subash Chandra Bose, Kamel Besbes, "Lane Departure Detection Using Image Processing Techniques," in *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*, 2017.
- [2] Byambaa Dorj and Deok Jin Lee, "A Precise Lane Detection Algorithm Based on Top View Image Transformation and Least-Square Approaches," in *Journal of Sensor*, Gunsan, Korea, 2015.
- [3] Ruyi Jiang, Mutsuhiro Terauchi, Reinhard Klette, Shigang Wang, Tobi Vaudrey, "Low-Level Image Processing for Lane Detection and Tracking," in *ArtsIT*, 2009.
- [4] M. Haloi and D. B. Jayagopi, "A robust lane detection and departure warning system," in *IEEE Intelligent Vehicles Symposium*, 2015.
- [5] Jihun KimMinho Lee, "Robust Lane Detection Based On Convolutional Neural Network and Random Sample Consensus," in *International Conference on Neural Information Processing*, 2014.
- [6] Marc Revilloud , Dominique Gruyer, Evangeline Pollard, "An improved approach for robust road marking detection and tracking applied to multi-lane estimation," in *2013 IEEE Intelligent Vehicles Symposium (IV)*, 2013.
- [7] K. Zhaom, M. Meuter, C. Nunn, D. Muller, S. Muller-Schneiders, and J. Pauli, "A novel multi-lane detection and tracking system," in *A novel multi-lane detection and tracking system*, 2012.
- [8] Yingmao Li, Asif Iqbal, Nicholas R. Gans, "Multiple lane boundary detection using a combination of low-level image features," in *IEEE 17th International Conference*, 2014.
- [9] Soomok Lee, Seong-Woo Kim, Seung-Woo Seo, "Accurate ego-lane recognition utilizing multiple road characteristics in a Bayesian network framework," in *IEEE Intelligent Vehicles Symposium*, 2015.
- [10] M. Aly, "Real time detection of lane markers in urban streets," in *IEEE Intelligent Vehicles Symposium*, 2008.
- [11] Karsten Behrendt and Jonas Witt, "Deep Learning Lane Marker Segmentation From Automatically Generated Labels," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.
- [12] D. Neven, B-D.Brabandere, S. Georgoulis, M. Proesmans and L-V. Gool, "Towards End-to-End Lane Detection: an Instance Segmentation Approach," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018 .
- [13] Thanh-Phat Nguyen, Vu-Hoang Tran, Ching-Chung Huang, "Lane Detection and Tracking Based on Fully Convolutional Networks and Probabilistic Graphical Models," in *IEEE*, 2018.
- [14] T. Emara, H. E. A. E. Munim and H. M. Abbas, "LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*, Perth, Australia, Australia, 2019.
- [15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.
- [16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2017.
- [17] S. Suzuki and a. others, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, p. 32-46, 1985.
- [18] Wikipedia, "Polynomial interpolation," [Online]. Available: [https://en.wikipedia.org/wiki/Polynomial\\_interpolation](https://en.wikipedia.org/wiki/Polynomial_interpolation).
- [19] G. Brostow, J. Fauqueur and a. R. Cipolla, "Semantic object classes in video: A high-definition ground truth database,," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88-97, 2009.
- [20] Kingma, D. P. a. Ba and J. L. Adam, "A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.
- [21] J. Long, E. Shelhamer and a. T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
- [22] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani and M. Jagersand, "RTSeg: Real-Time Semantic Segmentation Comparative Study," in *IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018.

[1] Jamel Baili, Mehrez Marzougui , Ameer Sboui , Samer Lahouar , Mounir Hergli , J.Subash Chandra Bose, Kamel Besbes, "Lane



# A Lightweight Model For Real-time Traffic Sign Recognition

Trung-Hieu Nguyen  
Faculty of Vehicle and Energy  
Engineering  
HCMC University of Technology  
Education  
Ho Chi Minh City, Viet Nam  
hieuntr@hcmute.edu.vn

Vu-Hoang Tran  
Faculty of Electrical and Electronics  
Engineering  
HCMC University of Technology  
Education  
Ho Chi Minh City, Viet Nam  
hoangtv@hcmute.edu.vn

Van-Dung Do  
Faculty of Vehicle and Energy  
Engineering  
HCMC University of Technology  
Education  
Ho Chi Minh City, Viet Nam  
dodzung@hcmute.edu.vn

Van-Thuyen Ngo  
Faculty of Electrical and Electronics Engineering  
HCMC University of Technology Education  
Ho Chi Minh City, Viet Nam  
thuyen.ngo@hcmute.edu.vn

Thanh-Thanh Ngo-Quang  
Faculty of Electrical and Electronics Engineering  
HCMC University of Technology Education  
Ho Chi Minh City, Viet Nam  
thanhnqt@hcmute.edu.vn

**Abstract**— Traffic sign detection (TSR) is a hot topic in the field of computer vision with lots of applications such as autonomous vehicles, path planning, robot navigation etc. Especially, it also can be applied in advanced driver assistance system (ADAS) which can help drivers get more useful information on the road to make decision exactly. However, most of the developed systems cannot be used in real-time environment. Therefore, in this paper, a light-weight model has been proposed for real-time traffic sign recognition. Our system is embedded on a 1:7 RC vehicle and tested on our small driving environment. The experimental results demonstrate the effectiveness and robustness of the proposed model in many challenging situations.

**Keywords**— Traffic sign recognition, light-weight model, ADAS, deep learning, SVM

## I. INTRODUCTION

Thanks to the rapid growth of deep learning, the advanced driver assistance systems (ADAS), as shown in Fig. 1, has been strongly exploited and continuously improved. They bring many benefits to drivers when traveling on the road. Among the features of ADAS, traffic sign recognition plays the key role and attracts a lot of attention recently. In real environment, in order to detect traffic sign, we will face to complex challenges such as lighting effects, blur and fade, motion artifact, chaotic backgrounds, and viewing angle problems, etc. By handling these problems, we are able to create a robust system which can detect, localize traffic signs and provide valuable information about the driving scenario.

Some related methods for TSR are summarized corresponding to the faced problems are shown in Table I. These methods have their own advantages and disadvantages. The conventional image processing methods are quite fast and can be used in embedded system, however their performances are limited due to the effects of some real challenges. Meanwhile, the deep learning based methods achieve the better performance however, they are usually quite heavy and cannot be used in real-time on the embedded system. Our aim is to incorporate the best of these methods to design the most appropriate system for our application. Each method can be leveraged in different steps depend on its properties. In general, the object detection problem consists of two main tasks: (1) identifying the location of the

considered objects and (2) recognizing them. For the first task, SSD [2] shows its strengths because it can solve most of the problems and also can detect many objects with different scales in one run. However, the original SSD model is quite heavy, while our final target is embedded our algorithm into NVIDIA Jetson TX2, which only can work smoothly with a lightweight model. To be able to operate on embedded systems, a modification in network structure is needed to reduce the number of weights of the model. For the second task, we realize that the Histogram of Oriented Gradients (HOG) features [4] is quite simple and sufficient for classifying the basic traffic signs. With HOG features, we can simply use an SVM model with a linear kernel, which is pretty fast and lightweight, as a classifier.

In summary, our proposed method takes advantage of the previous methods into our unified system to resolve the main problems in TSR. Our goal is to have a reliable and lightweight model that can work on NVIDIA Jetson TX2. The paper is organized as follows: Section II describes the proposed method. Section III shows the experimental results and comparison. In section IV, we conclude the paper.

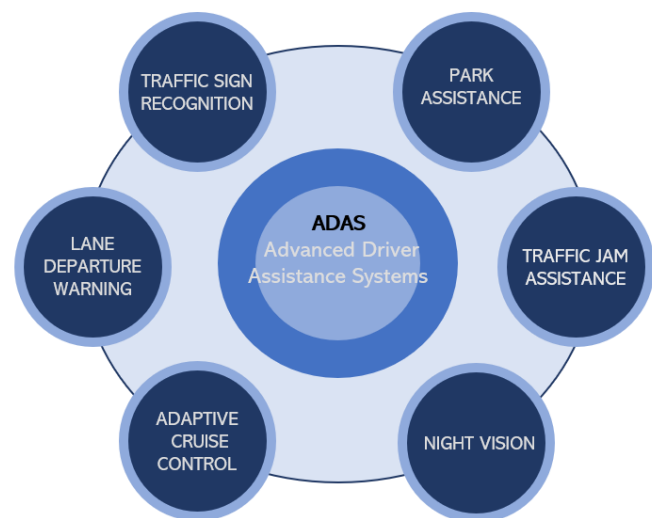


Fig 1. The Advanced Driver Assistance System.

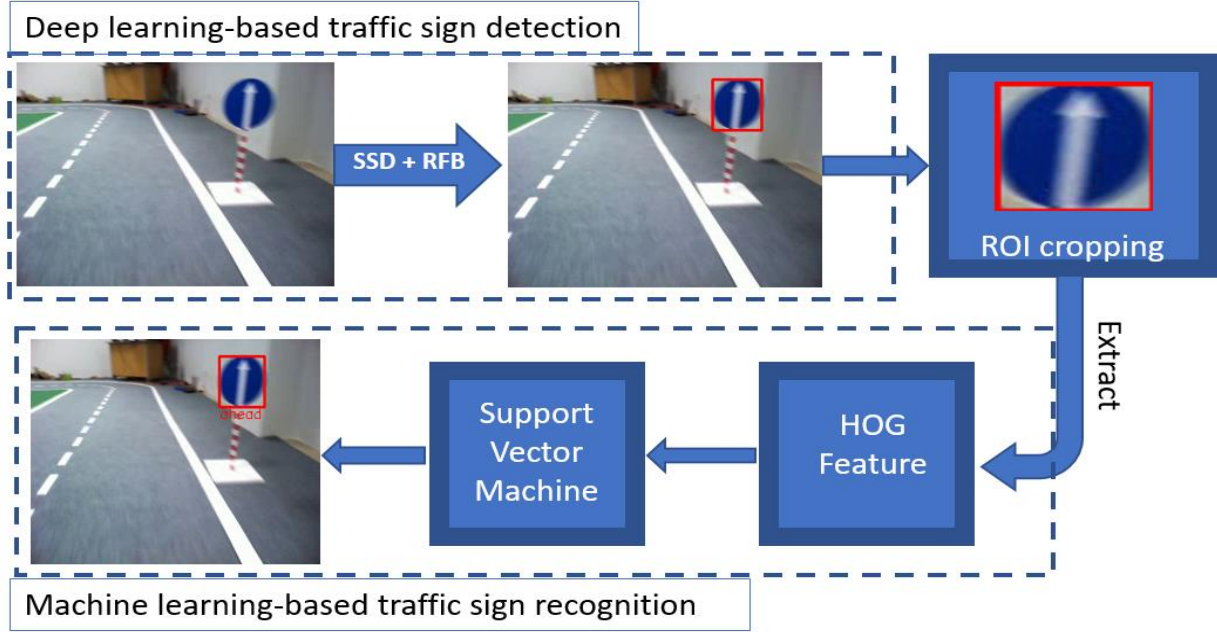


Fig 2. The proposed framework.

TABLE I. CHALLENGES AND METHODS ON TSR. LE: LIGHT EFFECTS. BF: BLUR AND FADE. MA: MOTION ARTIFACT. CB: CHAOTIC BACKGROUND. VA: VIEW ANGLE

Methods	LE	BF	MA	CB	VA	Real-time	Task	
							1	2
Image Processing		✓	✓			✓	✓	
SVM [1]	✓			✓	✓	✓		✓
SSD [2]	✓	✓	✓	✓	✓		✓	✓
Faster RCNN [3]	✓	✓	✓	✓	✓	✓	✓	✓

## II. PROPOSED METHOD

### A. The pipeline of the proposed framework

The details of our framework are described in Fig 2. It consists of 2 tasks: (1) identifying the location of the traffic signs and (2) recognizing them. As shown in Table 1, SSD is quite impressive when solving most of the problem in the first task. Therefore, in this paper, we propose to use the SSD-like structure, as shown in Fig. 3, to detect the location of traffic signs. In order to have a light weight model, we use mobilenet\_v1 network [5] as backbone. Besides, to enhance the feature discriminability and robustness of model, we replace some middle-level layers of mobilenet by one RFB layer [6]. After modifying based on the model in [7], our model now is reduced to only 1.14MB compared to the original SSD (20MB). After detecting the location, the traffic sign bounding boxes are cropped and send to the next phase, traffic sign recognition.

As shown in Table 1, SVM, a light-weight classifier, can handle this task in real-time on the embedded system. Therefore, in this paper, we propose to use SVM to identify the type of detected traffic signs. However, SVM is a simple classifier, in order to achieve the good performance, we need to ensure the extracted features are distinguishable. In this

paper, we use HOG [4] to extract features from detected traffic signs. In order to understand data structure, we visualize our data with HOG features into 2-D using t-SNE [8] as shown in Fig 4. As described in the figure, data of each category have been clearly clustered. So, it does not matter when we use the simple classifier such as SVM to classify them.

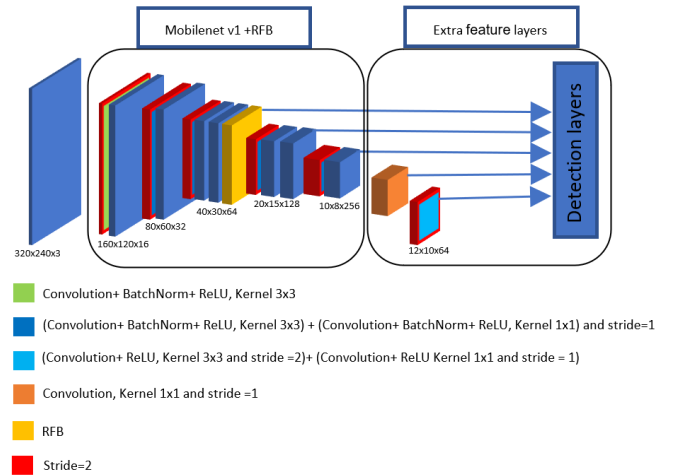


Fig 3. Architecture of the traffic signs recognition model.

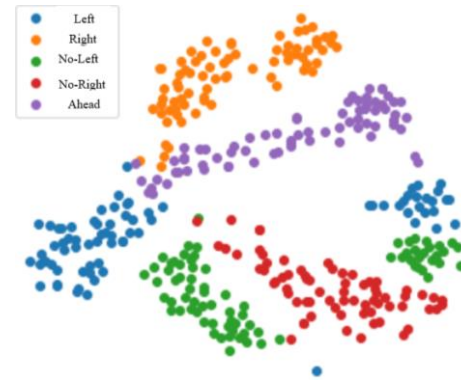


Fig 4. Visualizing HOG features in 2D using t-SNE.



### B. The loss function

For training the network of task 1, the loss of SSD method [2], is used. Assume that  $x_{ij}^k = \{1, 0\}$  be an indicator for matching the  $i^{th}$  default box ( $d$ ) to the  $j^{th}$  ground truth box ( $g$ ) of category  $k$ ;  $l$  denotes the predicted box;  $c$  represents the class confidence and  $N$  is the number of matched default boxes. The loss function is given as in Eq. 1.

$$L(x, c, l, g) = \frac{1}{N} (L_{conf}(x, c) + L_{loc}(x, l, g)). \quad (1)$$

Assume that the bounding box is represented by 4 parameters: center ( $cx, cy$ ), width ( $w$ ) and height ( $h$ ). The localization loss  $L_{loc}$  is given in Eq. 2.

$$\sum_{i \in Pos} \sum_{m \in \{cx, cy, w, h\}} \sum x_{ij}^k smooth_{L1}(l_i^m - \hat{g}_j^m), \quad (2)$$

where:

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx}) / d_i^w, \quad (3)$$

$$\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy}) / d_i^h, \quad (4)$$

$$\hat{g}_j^w = \log \left( \frac{g_j^w}{d_i^w} \right), \quad (5)$$

$$\hat{g}_j^h = \log \left( \frac{g_j^h}{d_i^h} \right). \quad (6)$$

The confidence loss  $L_{conf}$  is given in Eq. 7.

$$L_{conf}(x, c) = - \sum_{i \in Pos} x_{ij}^p \log(\hat{c}_i^p) - \sum_{i \in Neg} \log(\hat{c}_i^0), \quad (7)$$

$$\text{where, } \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}.$$

For task 2, we use the loss function of linear SVM and add the  $L_2$  regularization term as shown in Eq. 8. Where the raw model output is  $\theta^T x$ ,  $C$  is a fixed constant.

$$J(\theta) = C \sum_{i=0}^m [y^{(i)} Cost_1(\theta^T(x^{(i)})) + (1 - y^{(i)}) Cost_0(\theta^T(x^{(i)}))] + \frac{1}{2} \sum_{j=1}^n \theta_j^2, \quad (8)$$

where,  $\theta$  is the weight of the model;  $x$  and  $y$  are input HOG features and the corresponding label respectively;  $C$  is selected with a value of 0.005 to balance the different terms in loss function.

## III. EXPERIMENTAL RESULTS

### A. Datasets

In the experiments, we use the Kinect camera mounted on 1:7 RC vehicle, as shown in Fig. 5, to collect data in our small driving environment shown in Fig. 6. We first labeled 500 images collected from the camera. In order to increase the dataset's size, after cropping traffic sign regions (called ROI) in the images, we apply some augmentation methods such as brightness changing, shadow adding, noise adding, ect. Then we randomly collaged them into original images to obtain over 10,000 images of 5 different kinds of traffic sign as shown in Fig 7.

In order to comprehensively measure the accuracy of the proposed method, the precision and recall rates and confusion matrix are used. The precision and recall rates are based on three indicators: the false positive ( $FP$ ), the true

positive ( $TP$ ) and the false negative ( $FN$ ).  $TP$  and  $FP$  refer to the ratio of correctly and falsely detected objects in all region proposals.  $FN$  refers to the number of regions which include objects that should be detected but are not proposed. The definitions of precision and recall are as follows:

$$\text{Precision} = \frac{TP}{FP + TP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$



Fig. 5. 1:7 RC vehicle used in our experiments



Fig 6. Small driving environment.



Fig 7. Traffic-Sign Categories.

### B. Implemental Details

In training, before feeding data into our model in task 1, we use some image processing techniques such as photometric distortions (random brightness, contrast, hue, saturation, lighting noise), geometric distortions (random crop, mirror), and then resize them into dimensions of 320-by-240 pixels. The epochs and the batch size are relatively 100 and 24. For task 2, traffic sign recognition, all of the cropped traffic signs are stored in *png* format and resized into dimensions of 20-by-20 pixels.

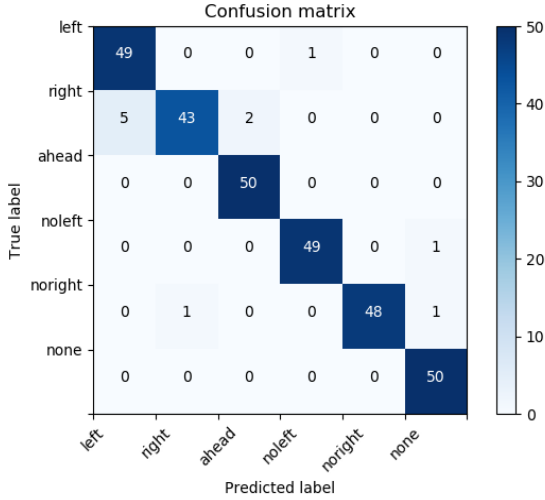


Fig 8. Recognition result on test sets

### C. Result

Our system can be used in real-time with the response time is 22 to 23 frame-per-second (fps) on NVIDIA Jetson TX2. In Table II, we show the fps for each task. To evaluate the performance, we collected data in real environment by running our car 5 times and treated them as test sets. The precision and recall rates of traffic sign detection system and the confusion matrix of recognition system are shown in Table III and Fig. 8 respectively. Based on the results, we can conclude that our method achieve the good detection performance on the embedded system with the average accuracy of 99.78%. Even when lighting conditions change, system performance is not affected as shown in Fig. 9. However, our system is still simple, it only considers 5 kinds of common traffic signs. To develop it in practice, we need to expand further.

TABLE II. FPS FOR EACH TASK

Task	Response time
Identifying the location of the considered objects	25 fps
Recognizing objects	30 fps
All	22~23 fps

TABLE III. DETECTION RESULTS IN DIFFERENT DATASETS

Dataset	FPR (%)	FNR (%)	ACC (%)
1	0.69	0.025	99.65
2	0.99	0.012	99.81
3	0.12	0.023	99.73
4	0.55	0.017	99.87
5	0.57	0.019	99.84
Average	<b>0.58</b>	<b>0.019</b>	<b>99.78</b>

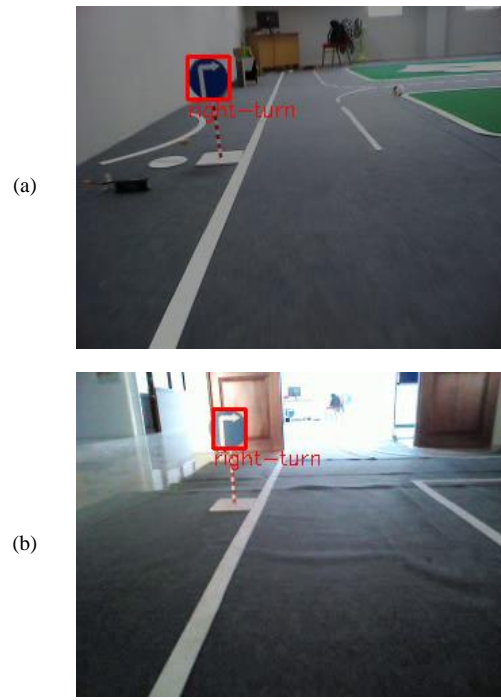


Fig 9. (a) Low lighting condition. (b) High lighting condition

### IV. CONCLUSION

This paper we proposed a light-weight model for traffic signs recognition. We take advantage of the previous methods into our unified system to achieve a reliable and lightweight model that can work on NVIDIA Jetson TX2. To detect the traffic signs location, we propose to use SSD-like structure with the combination between mobilenet\_v1 and RFB. For recognition task, the duo of HOG features and SVM is used. The obtained results show that our system work well in our tesing environment with a 1:7 RC vehicle. In the future, we will try to embed the system in a real vehicle and consider more kinds of traffic signs.

### REFERENCES

- [1] Ashanira Mat Deris, Azlan Mohd Zain, Roselina Sallehuddin, "Overview of Support Vector Machine in Modeling Machining Performances," in *Procedia Engineering*, 2011.
- [2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, 14th European Conference, 2016, pp. 21-37.
- [3] Chenchen Ji, Mingfeng Lu, Jinmin Wu, Zhen Guo, "Faster region-based convolutional neural network method for estimating parameters from Newton's rings," in *Modeling Aspects in Optical Metrology VII*, 2019.
- [4] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE*, 2005.
- [5] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," in *IEEE/CVF*, 2017.
- [6] Songtao Liu, di Huang, Yunhong Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," 2017.
- [7] Linzai, "Github.com," [Online]. Available: <https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB>.
- [8] Laurens van der Maaten, Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

# Design of Delta Robot Arm based on Topology Optimization and Generative Design Method

Thanh Hai Tuan Tran  
Dept. Machine Design and Industrial  
Systems Engineering  
The University of Danang, University of  
Science and Technology  
Danang, Vietnam  
tthtuan@dut.udn.vn

Dinh Son Nguyen\*  
Dept. Engineering Mechanics  
The University of Danang, University of  
Science and Technology  
Danang, Vietnam  
ndson@dut.udn.vn

Nhu Thanh Vo  
Dept. Mechatronics  
The University of Danang, University of  
Science and Technology  
Danang, Vietnam  
vnthanh@dut.udn.vn

Hoai Nam Le  
Dept. Mechatronics  
The University of Danang, University of Science and Technology  
Danang, Vietnam  
lehoainam@dut.udn.vn

**Abstract** — Delta robot is a kind of parallel robot that is widely used in industrial applications for its impressive advantages compared to the serial robot such as fast-moving speed and great productivity assembly. It is very useful for classification products and assembly requiring high processing speed. Saving materials and costs during the design and manufacturing process while ensuring the working capabilities of robots such as stiffness and durability is very important. Therefore, the authors propose an approach of utilizing the topology optimization method in the design process to solve such a problem. In this paper, the process of studying and applying the topology optimization method to design the Delta robot arm is presented. The redesigned product has a lighter weight, consumes less material, and therefore has lower manufacturing costs but still ensures the mechanical properties and the working ability requirement.

**Keywords**— *Topology optimization; Generative Design; Delta robot; Parallel robot*

## I. INTRODUCTION

Nowadays, Delta robot is widely used in many domains, from medical, military to industrial production. It is known as the robot to pick up and drop out products at a very fast speed. This kind of parallel robot which is invented by Reymond Clavel [1] with the outstanding advantages in the comparison with the serial robot such as high rigidity and load capacity, etc.

In the current competitive and globalization context, a short time to market plays an important role in determining the success of a product. In particular, the time spent designing the product takes up a large amount of time to complete the final one. Besides, manufacturing technologies, materials, and costs are constraints hindering freedom in the product design. Thus, it is necessary to have a useful tool to help designers in solving these problems.

Indeed, designers can make different numerical models of the designed product on the computer thanks to the remarkable advancement of the computer-aided-design (CAD) technology. Especially, generative design is a capability of CAD applications in which designers use specific algorithms to generate several design alternatives based on a set number of constraints. In addition, the appearance of additive manufacturing technology has solved major obstacles in the production of products with high geometric complexity. Therefore, the generative design is increasingly becoming a

helpful tool in design support and shortening design time compared to the traditional design one. The generative design supports key capabilities like topology optimization helps designers to find the optimal design with the least amount of material while ensuring other constraints.

Topology optimization is a method including algorithms that allows optimizing the distribution of the materials in a design space based on the loading, material, and rigidity constraints of the product. The remarkable advantage of this method is that product designers can design any shape in the design space. Besides, thanks to the reduction of material, the products with lighter weight can be created. Hence, the cost of material in the manufacturing process can be saved. Moreover, the less usage of material, the more ecological in industrial production.

There are some researches about topology optimization [2-4] and applications of this method in practical problems such as design parts in aeronautical engineering [5-6] or a technical part like an open-end wrench [7]. In robot design, the topology optimization method has been applied in humanoids [8-9].

At this moment, in the Faculty of Mechanical Engineering, the third generation of Delta robot is under development with the aim to apply in the industrial process. Thus, this paper proposes a design methodology using a generative design and topology optimization approach to design the upper arm of Delta robot reducing the weight, saving the cost of manufacturing but still maintain the mechanical properties and working abilities.

## II. METHODOLOGY

The paper proposes a procedure that allows designing the upper arm of the Delta robot using the topology optimization method. The procedure given in Fig. 1 includes the following steps.

### • Step 1: Determining input requirements

It is necessary to identify the input constraints of the Delta robot that we need to design. Then, the product designers should analyze constraints applying on the upper arm such as functional surfaces, load systems.

\* Corresponding author

- *Step 2: Classifying design and non-design space*

After determining the constraints applying on the upper arm of the robot, we need to design the geometric model of the arm. Then, it needs to classify design and non-design space. The design space of the arm is a space that the materials can be reduced in the optimization process.

- *Step 3: Discretizing spaces using the finite element method*

Spaces including design and non-design space need to be discretized by using the finite element method. The results are used in the next step to find the optimized solution.

- *Step 4: Finding the optimal solution*

The topology optimization algorithm is used to solve constraints from step 3. Some materials in the design space of the arm are removed in order to reduce the weight while ensuring its stiffness.

- *Step 5: Reviewing criteria*

From the result of optimization, the designers should evaluate the constraints of displacement, stress, etc. of the upper arm.

- *Step 6: Redesigning*

The model of the arm from the topology optimization is not a final design because of the roughness of surfaces. Thus, it needs to smooth surfaces of this numerical model to obtain the final design.

After these steps, we will get a new design of the upper arm with lighter weight but still guarantee the ability to work as required.

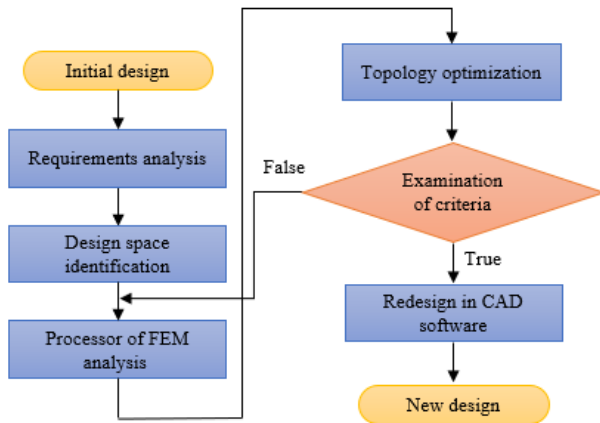


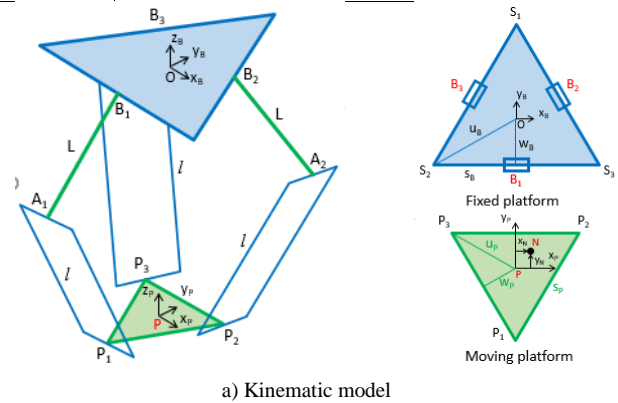
Fig. 1. The process to design a product using topology optimization method.

### III. CASE STUDY

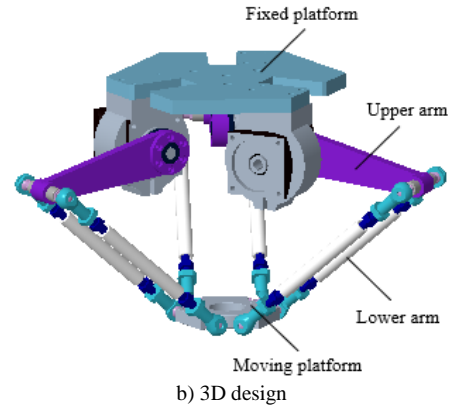
The arm design of the Delta robot in the laboratory of the Faculty of Mechanical Engineering will be proposed using the topology optimization method to reduce material consumption in the design process. The kinematic model and 3D design of the Delta robot are illustrated in Fig. 2 [10]. The definitions of geometric parameters are shown in Table 1.

TABLE I. DEFINITIONS OF GEOMETRIC PARAMETERS OF THE DELTA ROBOT

Parameter	Definition
$P_i$	The connection points between lower arms and moving platform ( $i = 1, 2, 3$ )
$s_B$	Length of the edges of fixed platform
$w_B$	Distance between $O$ and the edges of fixed platform
$u_B$	Distance between $O$ and the vertices of fixed platform
$s_P$	Length of the edges of moving platform
$w_P$	Distance between $P$ and the edges of moving platform
$u_P$	Distance between $P$ and the vertices $P_i$ ( $i = 1, 2, 3$ ) of moving platform
$L$	Length of the upper arm $B_iA_i$ ( $i = 1, 2, 3$ )
$l$	Length of lower arms
$h$	Width of lower arms



a) Kinematic model



b) 3D design

Fig. 2. Model of Delta robot.

The Fig. 3 shows the model of the upper arm in CAD software.



Fig. 3. Geometric model of the upper arm.



### A. Requirements analysis

In this part, the object separation method is used to release the connections between parts in the Delta robot. From there, we can determine the forces acting on the part under consideration. The analysis of the Delta robot dynamics in three-dimensional space is very complicated. For simplicity, we use the Delta robot model in the 2D plane (see in Fig. 4).

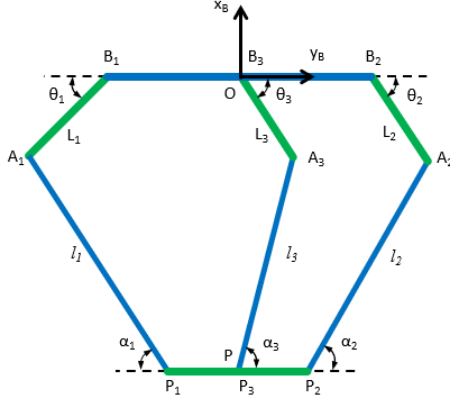


Fig. 4. Static analysis diagram of the robot.

Where:

$B_1B_2$ : Distance between the centers of revolute joints  $B_1$  and  $B_2$

$L_1 = L_2 = L_3 = L = 0,2m$ : Length of the upper arms

$l_1 = l_2 = l_3 = l = 0,28m$ : Length of the lower arms

$s_B = 0,1\sqrt{3}m$ : Length of the edges of fixed platform

$s_P = 0,05\sqrt{3}m$ : Length of the edges of moving platform

$m_{l_1} = m_{l_2} = m_{l_3} = m = 0,2kg$ : Weight of lower arms

The force acting on each robot arm is analyzed separately. Each part of the arm is considered a separate object. The acting forces on each part of the arm include moving platform, the upper arm, and lower arm are indicated in Fig. 5.

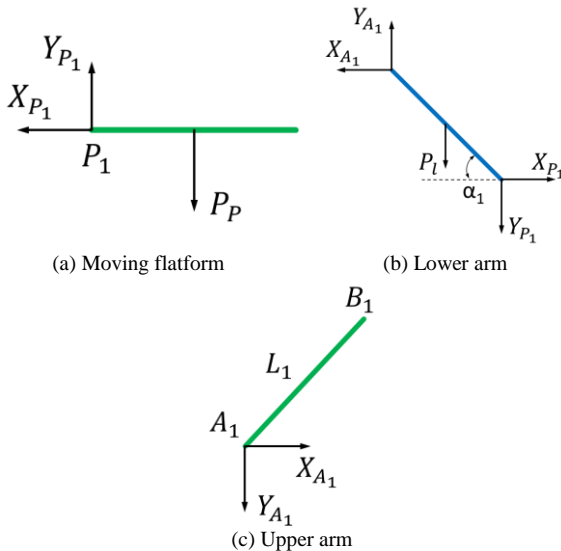


Fig. 5. Force systems acting on the parts of the arm.

Applying the equilibrium force equations for each part, we obtain the systems of equations (1), (2).

$$\begin{cases} X_{P_1} = 0 \\ Y_{P_1} = P_P \end{cases} \quad (1)$$

$$\begin{cases} Y_{A_1} = P_{l_1} + Y_{P_1} \\ X_{A_1} = \frac{(P_{l_1} + Y_{P_1}) \cos \alpha_1 - P_{l_1} \frac{\cos \alpha_1}{2}}{\sin \alpha_1} = \frac{\frac{P_{l_1}}{2} + Y_{P_1}}{\tan \alpha_1} \end{cases} \quad (2)$$

To facilitate the setting forces acting on the upper arm in the optimization software, we transform its coordinate system to a beam system with a fixed support at one end (point  $B_1$ ) and a freedom head (point  $A_1$ ), as shown in Fig. 6.

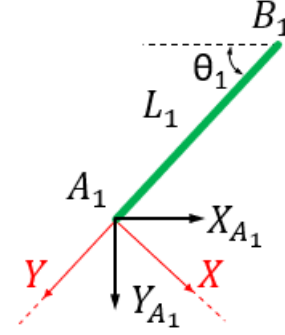


Fig. 6. Transformed coordinate system applying on the upper arm.

Where:

$\vec{X}_{A_1}, \vec{Y}_{A_1}$  are initial forces acting on the upper arm

$\vec{X}, \vec{Y}$  are transformed forces acting on the upper arm

Then, we have the value of  $\vec{X}, \vec{Y}$  as follow:

$$\begin{cases} X = \frac{1}{2} \left( \frac{P_{l_1} + Y_{P_1}}{\sin \theta_1} - \frac{\frac{P_{l_1}}{2} + Y_{P_1}}{\tan \alpha_1 \cos \theta_1} \right) \\ Y = \frac{1}{2} \left( \frac{P_{l_1} + Y_{P_1}}{\cos \theta_1} + \frac{\frac{P_{l_1}}{2} + Y_{P_1}}{\tan \alpha_1 \sin \theta_1} \right) \end{cases} \quad (3)$$

From the equations (1), (2), (3), we extract the angles  $\theta_1$  and  $\alpha_1$  to maximize  $|\vec{X}|, |\vec{Y}|$ . By using the vector method, we have the equations (4).

$$\begin{cases} -l \cos \alpha_1 + L \cos \theta_1 + \frac{s_B}{2} - \frac{s_P}{2} = 0 \\ l \sin \alpha_1 + L \sin \theta_1 = 0 \end{cases} \quad (4)$$

This is the maximum optimization problem with the objective function (5).

$$\begin{cases} \max f(\theta_1, \alpha_1) = \frac{1}{2} \left( \frac{P_{l_1} + Y_{P_1}}{\sin \theta_1} - \frac{\frac{P_{l_1}}{2} + Y_{P_1}}{\tan \alpha_1 \cos \theta_1} \right) \\ \max f(\theta_1, \alpha_1) = \frac{1}{2} \left( \frac{P_{l_1} + Y_{P_1}}{\cos \theta_1} + \frac{\frac{P_{l_1}}{2} + Y_{P_1}}{\tan \alpha_1 \sin \theta_1} \right) \end{cases} \quad (5)$$

The result of this problem is  $\theta_1 = -15^\circ$  và  $\alpha_1 = 75^\circ$ . In this position, the forces acting on the upper arm is maximum.

### B. Topology optimization procedure

The steps to perform topology optimization procedure include:

#### 1. Determining input requirements



The input requirements of three upper arms of the Delta robot in the laboratory include the lightweight, assembly constraints, and rigidity. The robot should be ensured to carry the maximum load of 2 kg. From these requirements, the forces applying on the arm of the robot are determined in the previous section. Then the design space, forces, and supports need to be defined in the software (see in Fig. 7).



Fig. 7. Model of design space, support and forces.

Moreover, we have to define the type of material which is used for the designed part. It depends on the method we use to manufacture the part.

## 2. Defining other constraints

Besides the load and material requirements of the design part, we need to define additional important constraints such as the connections of the design part with other parts, the definition of transposition, etc.

## 3. Classifying design and non-design space

The initial model of the upper arm includes three small solid units. To apply the topology optimization method, we need to define the design space where the usage of the material is reduced in the manufacturing process and non-design space which is fixed and does not participate in the optimization process. Besides, we also need to define the functional surfaces of the upper arm. They are the surfaces contacting the shaft of the motor and the lower arm as shown in Fig. 8.

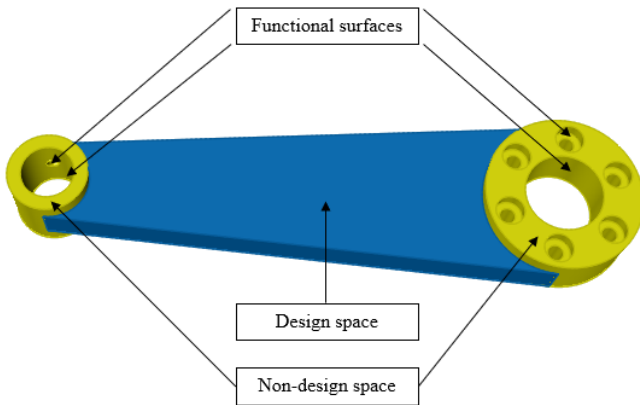


Fig. 8. Classification of design and functional spaces.

## 4. Discretizing spaces using the finite element method

All mesh elements are generated using the FEM analysis to discretize the design space. This is the preparation for topology optimization.

## 5. Performing topology optimization

In this case, we use Altair SolidThinking Inspire software to solve the optimization problems of the upper arm of the robot. The results of topology optimization are obtained with different types of input constraints. Some topology optimization solutions are given in Fig. 9. The result with 54% weight reduction is the best solution.

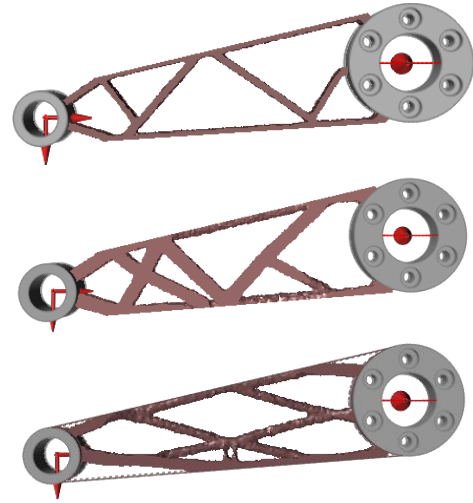


Fig. 9. Topology optimization results.

## IV. DISCUSSION

Based on the constraints of the workspace of the robot and its mechanical properties, we select a suitable model among the topology optimization results. The workspace of the robot is mainly depended on the values of the angles  $\theta_1$  and  $\alpha_1$ . For choosing values of  $\theta_1$  and  $\alpha_1$ , we select the highest value of  $\theta_1$  and  $\alpha_1$  that both satisfied the conditions of maximum material reduction as indicated in function (5) and the workspace requirement. After choosing a model corresponding to the satisfactory value of angles in the workspace of the robot, we have to concern the mechanical properties of this model. At the final step, we will try to choose the model which has the values of mechanical properties closest to the ones of the initial model. Table 2 shows a comparison between the properties with different safety factors.

TABLE II. MECHANICAL PROPERTIES WITH DIFFERENT SAFETY FACTORS

Safety Factors	1.2	1.3	1.4	1.5	Initial model
Properties					
Max Displacement (mm)	0.2544	0.2269	0.2224	0.2163	0.0710
Max Von Mises Stress (MPa)	60680	42180	62080	43950	11810
Max Shear Stress (MPa)	32820	23400	32850	22430	6256
Major Principal Stress (MPa)	83380	57210	53780	59770	10140
Mass Total (kg)	0.1947	0.2089	0.2065	0.2072	0.4430

We chose the result of optimization with the safety factor 1.5 because this factor has the mechanical properties that are closest to the initial model. The final model of the upper arm in the topology optimization procedure is transformed into CAD format. However, the quality of surfaces is not satisfied so we cannot directly use it to manufacture. Thus, it is mandatory to redesign a new model of the upper arm in CAD software based on the transformed model. The PolyNURBS tool in Altair SolidThinking Inspire software is used to redraw. In order to validate the design, requirements such as

mechanical properties, stiffness need to be analyzed and tested. The results of the FEM analysis are shown in Fig. 10.

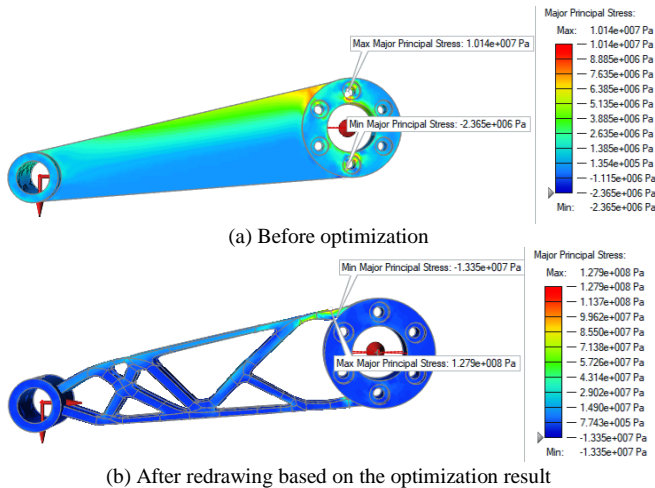


Fig. 10. Comparison of major principal stress of the upper arm Before optimization (a) and After redrawing based on the optimization result (b).

As indicated in Fig.10, the max value of Major Principal Stress in the upper arm after redrawing based on the optimization result is 10 times bigger compared to the initial upper arm design. Besides, the position of this value is located at the corner, and it is dangerous for the upper arm during its operation. Thus, we have to increase the radius of this corner to reduce stress, as shown in Fig. 11.

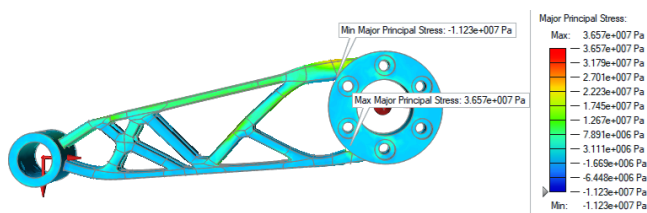


Fig. 11. Modification of the optimization result.

The Mass Total of the upper arm after redesigning is 0.227 kg and its weight is reduced by 48.8%.

If we use traditional manufacturing technologies such as milling or casting, it is impossible to fabricate the arm because of the complicated geometry. Thanks to the FDM additive manufacturing technology, the prototype of the upper arm of the robot in plastic material is fabricated (see in Fig. 12).

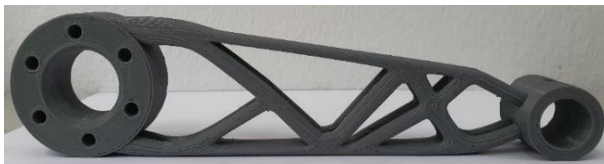


Fig. 12. The new upper arm manufactured by 3D printer.

## V. CONCLUSION

This study offered a design procedure using topology optimization and generative approaches to create a design solution space. The product designers can choose the design solution meeting fully requirements based on many solutions proposed by generative design. An innovative design of the upper arms of the Delta robot in the laboratory is presented in the paper to demonstrate the applicability of the proposed method in many different industrial applications. The reduction of material in products contributes to cost savings and environmental protection. In addition, the combination of topology optimization and additive manufacturing also plays a crucial role to shorten the time to bring the product to market.

## REFERENCES

- [1] R. Clavel, "DELTA, A fast robot with parallel geometry," 18th International Symposium on Industrial Robot, pp. 91-100, 1988.
- [2] M. P. Bendsoe and O. Sigmund, *Topological Optimization, Theory, Methods and Application*. Springer Verlag, Berlin, 2003.
- [3] O. Sigmund and S. Torquato, "Design of materials with extreme thermal expansion using a three-phase topology optimization method," *Journal of the Mechanics and Physics of Solids*, vol. 45, no. 6, pp. 1037-1067, 1997.
- [4] X. Y. Yang, Y. M. Xie and G. P. Steven, "Evolutionary methods for topology optimisation of continuous structures with design dependent loads," *Computers & Structures*, vol. 83, no. 12, pp. 956-963, 2005.
- [5] M. Süß, C. Schöne, R. Stelzer, B. Kloeden, A. Kirchner, T. Weißgärber and B. Kieback, "Aerospace Case Study on Topology Optimization for Additive Manufacturing," *Fraunhofer Direct Digital Manufacturing Conference DDMC*, pp. 37– 41, 2016.
- [6] A. W. Gebisa and H. G. Lemu, "A case study on topology optimized design for additive manufacturing," *IOP Conference Series: Materials Science and Engineering*, vol. 276, no. 1, pp. 012026, 2017.
- [7] D. S. Nguyen and F. Vignat, "Topology Optimization as an Innovative Design Method for Additive Manufacturing," *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 304-308, 2017.
- [8] W. Kwon, H. K. Kim, J. K. Park, C. H. Roh, J. Lee, J. Park, W-K. Kim and K. Roh, "Biped humanoid robot Mahru III," *7th IEEE-RAS International Conference on humanoid robots*, pp. 583–588, 2007.
- [9] S. Lohmeier, T. Buschmann and H. Ulbrich, "Humanoid robot LOLA," *IEEE International Conference on Robotics and Automation*, pp. 775–780, 2009.
- [10] H. N. Le and X. H. Le, "Geometrical Design of a RUU Type Delta Robot Based on the Prescribed Workspace," *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, pp. 359-364, 2018.

# Effects of Soaking Process on CBR Behavior of Geotextile Reinforced Clay with Sand Cushion

Tu Nguyen Thanh  
Faculty of Civil Engineering  
Ho Chi Minh University of Technology and Education  
Ho Chi Minh City, Vietnam  
tunt@hcmute.edu.vn

Duc Nguyen Minh  
Faculty of Civil Engineering  
Ho Chi Minh University of Technology and Education  
Ho Chi Minh City, Vietnam  
ducnm@hcmute.edu.vn

**Abstract**—Clay, which was excavated from the river, was difficult to reuse because of the massive property changes when changing its water content. When being saturated, the clay becomes looser and softer, inducing a significant reduction in the bearing capacity. To improve those disadvantages, the clay was reinforced by the nonwoven geotextile with a sandwich sand layer. Using the California Bearing Ratio (CBR) tests, the reinforced clay's bearing capacity behavior with a sand cushion under soaking condition was investigated. The result reveals that the sandwich sand layer significantly improved the CBR value of the reinforced riverbed clay. After 96 hours of soaking, the CBR value of reinforced specimens was as high as 1.5-2.8 times that of the un-reinforced specimen. Regarding the bearing capacity reduction after soaking, the CBR value of unreinforced riverbed clay was less than 3%, which reduced up to 73.1% of its bearing capacity before soaking. In contrast, the CBR reductions of reinforced specimens were varied from 42.2-60.8% depending on the thickness of the sand layer. When increasing the sand height, the CBR value went up, especially for soaking specimens. The optimal dry mass ratio between sand and soil was 0.1 in other that the CBR got the highest value.

**Keywords**—soaking, soft clay, swelling, CBR

## I. INTRODUCTION

Using riverbed clay instead of sand for backfill, especially in transportation construction like roads has many benefits: (1) not losing local cultivated land; (2) increasing the depth of river; (3) ensuring the elevation of roads adapted to the increases of water level due to global climate change and (4) green and solution for sustainable development. Nevertheless, there are some disadvantages: low shear strength, high void ratio, impermeability, and the massive change of properties when being soaked (after rainfall) [1-2]. Using soft clay as a backfill required a drainage system and construction method to ensure its strength [3-6]. Geotextile and sand cushion are usually used to enhance the strength of soil as well as handle weakness. The high permeability of geotextile significantly increases the bearing capacity and stability of reinforced soil structure [7]. Using geogrid-reinforced sand cushion increased the capacity of soft soil, and the subgrade reaction coefficient K30 was improved by 3000% as well as the deformation is reduced by 44% [8]. Reference [9] introduced the construction of a 3 m high embankment on the geocell foundation over the soft settled red mud, a waste product from the Bayer process of the Aluminum industry. In this case, the combination of geocell and geogrid was recommended to stabilize the embankment base. Reference [10] applied sand cushion combining with geotextile under breakwater on soft ground to constrain the lateral displacement of both the embankment and the ground, and the reinforcement suppressed the range of high-stress level in the system. In general, the weaker the ground is, the higher the modulus of the geotextile is, the more

effective the reinforcement would be. The geotextile and sand cushion could improve the bearing capacity of the reinforced soil by up to 7 times [11]. The important drainage role of geotextile in enhancing the bearing capacity and stability of soft soil in embankment constructions was also reported previously [7]. Encapsulating geogrids in thin layers of sand to enhance the strength of clay was investigated in the direct shear test [12], pullout tests [3, 13, 14], and triaxial compression test [15]. These results showed that a thin sand cushion improves the interface friction between clay and geotextile, increasing the strength of clay. This sand cushion was also a drainage boundary, decreasing the pore pressure in increasing loads. References [12, 13, 14, 15] showed the optimum height of sand was 8-10 mm in the unconsolidated - un drainage test (UU) and direct shear test, or even up to 8 cm in the pullout test. Regarding the drainage boundary, geotextile prevented the interlocking effect of fine particles of clay penetrated into the sand cushion layer [16, 17]. Geotextile also improved the bearing capacity of reinforced expansive clay up to 1.5 times for unsoaked condition and 3.3 times for soaked cases [18].

Many researchers performed the laboratory test to investigate the California Bearing Ratio (CBR) of reinforced soil. The CBR values of sand reinforced by high-density polyethylene (HDPE) increased up to 3 times [19]. Similarly, The CBR of clay reinforced with geogrid in soaked condition could be improved by 1.9 – 2.6 times [20]. For un-soaked specimens, the value of CBR was about 1.9-4.5 times that of unreinforced clay. The CBR enhancement of lateritic soil reinforced with one and two layers of geogrid was also observed. The higher number of reinforcement layers, the high the bearing capacity of reinforced specimens was [21].

Although there were many CBR tests to investigate the behavior of reinforced clay, the shear strength reduction of reinforced clay with sand cushion due to the soaking process was not fully determined. In this paper, a series of laboratory tests for CBR was performed to examine the bearing capacity of the soft clay reinforced by two non-geotextile layers covered a thin sand cushion layer. The CBR behavior of the reinforced specimens under soaked and unsoaked conditions was determined to quantify the bearing capacity reduction of specimens due to the soaking process.

## II. TEST MATERIALS

### A. Soft clay

Fig. 1 highlighted the grain-size distribution of riverbed clay based on reference [22]. The clay soil was the same as the clay in the previous study [18]. It was excavated from the Cai Lon River, Kien Giang province, with the water content,  $\omega = 57.4\%$ , and the void ratio,  $e = 1.6$ . The plasticity index, plastic

limit, and liquid limit are 46.6, 44.9, and 91.5, respectively. Using the Proctor compaction test [23], the optimum water content ( $w_{opt}$ ) is 26.6% with its maximum dry unit weight  $\gamma_{d,max} = 14.56 \text{ kN/m}^3$ . The clay is classified as high plastic inorganic silt (MH) according to the Unified Soil Classification System. Using the hydraulic conductivity of the clay,  $k_{sat}$  is  $= 1.18 \times 10^{-10} \text{ m/s}$  from one-dimensional consolidation test results.

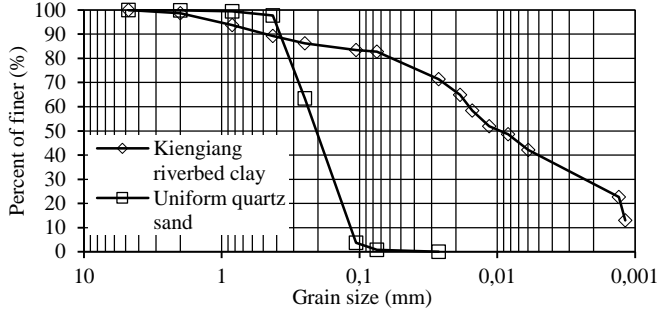


Figure 1. Grain-size distribution of soft soil and sand

### B. Geotextile

The geotextile used in the research is the same as the reinforcement material in the previous research [18], which was a commercially available needle-punched Polyethylene terephthalate (PET) nonwoven geotextile. Its cross-plane permeability ( $3.5 \times 10^{-3} \text{ m/s}$ ) is suitable for the lab test with  $1.96 \text{ s}^{-1}$  of the permittivity. The mass and thickness of PET are  $200 \text{ g/m}^2$  and  $2.78 \text{ mm}$ , respectively. The apparent opening size is  $0.11 \text{ mm}$ . Regarding the wide-width tensile test in the transverse direction, the PET gained  $9.28 \text{ kN/m}$  of ultimate tensile strength at  $84.1\%$  failure strain. While in the longitudinal direction, the ultimate tensile strength and the failure strain are  $7.08 \text{ kN/m}$  and  $117.8\%$ , respectively.

### C. Uniform quart sand

Table 1 presented the properties of used sand. Sand is classified as clean sand, few fine particles, poor gradation. The gain-size distribution of sand is shown in Fig 1.

TABLE I. SAND PROPERTIES

Property	Value
Unified Soil Classification System	SP
Specific gravity, $G_s$	2.66
$D_{10}$ (mm)	0.121
$D_{30}$ (mm)	0.169
$D_{60}$ (mm)	0.242
Coefficient of curvature, $C_c$	0.98
Coefficient of uniformity, $C_u$	2.00
Minimum dry unit weight, $\gamma_{d,min}$ (kN/m <sup>3</sup> )	12.56
Maximum dry unit weight, $\gamma_{d,max}$ (kN/m <sup>3</sup> )	15.43
<b>At relative density, <math>D_r = 0.9</math></b>	
Dry unit weight, $\gamma_d$ (kN/m <sup>3</sup> )	15.09
Friction angle from direct shear test, $\phi'$ (deg)	35.1
Geotextile/sand interface friction angle, $\phi'_a$ (deg)	23.7
Efficiency factor, $E = \tan \phi'_a / \tan \phi'$	0.62

## III. EXPERIMENTAL PROGRAM

There were total 10 specimens with the variation of soaking conditions, and the thickness of the sand cushion layer, which changed from 0 cm (unreinforced) to 4 cm (Fig. 2).

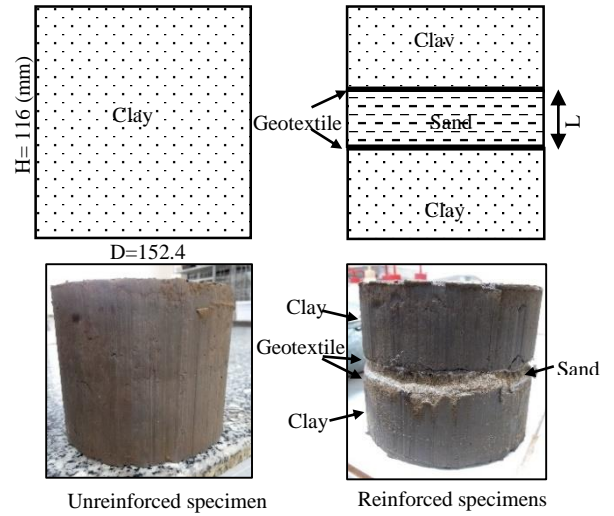


Figure 2. Geotextile and sand cushion arrangement in reinforced and unreinforced specimens

### A. Specimen preparation

To prepare soil specimens, a natural clay was excavated from the riverbed in the form of wet bulk. It was placed in an oven (temperature was set at less than  $60^\circ \text{C}$  for a minimum of 24 hours and then crushed and grounded into a dry powder in a mortar. After mixing different quantities of powder and water corresponding to the desired water content, specimens were placed in a plastic bag in a temperature-controlled chamber for a minimum of 2 days to ensure a uniform distribution of water in the soil mass.

For un-reinforcement specimens, a mold with 116 mm height and a diameter of 152.4 mm was used to prepare the specimens by 5 compaction layers. Each soil layer was compacted by 10 blows/layer (equivalent to  $482 \text{ kJ/m}^3$  of compaction energy) at the optimum water content, which was found by several trial compaction tests (Fig 3).

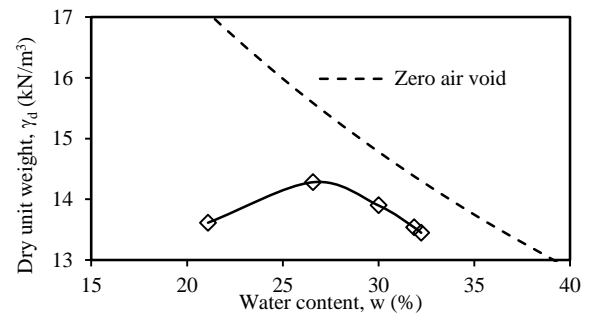


Figure 3. Compaction behavior of the clay under a modified compaction energy,  $E = 482 \text{ kJ/m}^3$

For specimens reinforced by geotextile and sand cushion, after each soil layer was compacted and leveled, the soil surface was scarified before two 15.24 (mm)-diameter dry geotextiles layers were placed horizontally on the roughed surface of soil and sand. The sand was compacted to reach  $15.09 \text{ kN/m}^3$ , which equivalent to 90% of relative density (Table 1).

For the soaked specimens, the compacted specimens were soaked in 96 hours before performing the CBR test. The surface of specimens was loaded using a surcharge of 4.54 kg mass. A 2.27 kg weight was placed to prevent the upheave of soil into the hole of surcharge. During the soaking process, the swell of specimens was recorded frequently after every 1-2 hours.

#### B. CBR testing

Based on the reference [24], the rate of penetration is approximately 0.05 inches/min (1.27 mm/min) in the CBR laboratory test. The stress in the piston was recorded with time and corrected due to the surface irregularities or other causes, as recommended by [24]. The value of CBR was determined as follows:

$$\text{CBR}_1 (\%) = P_1 / 6.9 \times 100 \quad (1)$$

$$\text{CBR}_2 (\%) = P_2 / 10.3 \times 100 \quad (2)$$

in which  $\text{CBR}_1$  and  $\text{CBR}_2$ : the CBR value at 2.54 mm and 5.09 mm of penetration, respectively;  $P_1$ ;  $P_2$ : the value of corrected stress in piston (MPa) at 2.54 mm and 5.09 mm, respectively.

If  $\text{CBR}_1 \geq \text{CBR}_2$ , the CBR is  $\text{CBR}_1$ . Otherwise,  $\text{CBR}_1 < \text{CBR}_2$ , do the test again and if the results are the same, use the  $\text{CBR}_2$  as the CBR value.

### IV. RESULTS AND DISCUSSION

#### A. Influence of nonwoven geotextile and sand cushion on the swell behavior

The swell of the specimen ( $S$ ) is considered the swell of soil only. It is defined as the ratio between swell and the original height of specimen in percent as follows.

$$S = s / H_{\text{soil}} \quad (3)$$

in which  $s$  is vertical swell measured with time;  $H_{\text{soil}}$  is the height of soil only (exclude the thickness of reinforcement layers if any) before soaking.

The percent swell of unreinforced and reinforced specimens ( $S$ ) in time is given in Fig. 4. Generally, it increased by the time during the soaking process. The swell of the specimens reached the equilibrium after 96 h of soaking.

At the first of 30 hours, the percent swell of reinforced specimens is higher than that of unreinforced specimens (Fig. 4a). However, at the end of the soaking process, the swells of reinforced specimens were slightly smaller than that of the unreinforced specimen (Table 2). The effect is due to the local lateral confinement from soil-reinforcement interaction. It can be explained that the expansion develops in all directions and mobilizes the interfacial frictional force between soil and reinforcement [19]. This frictional force tends to counteract the swelling pressure in a direction that parallels the reinforcement and consequently reduces the heave. A similar observation was found by reference [25].

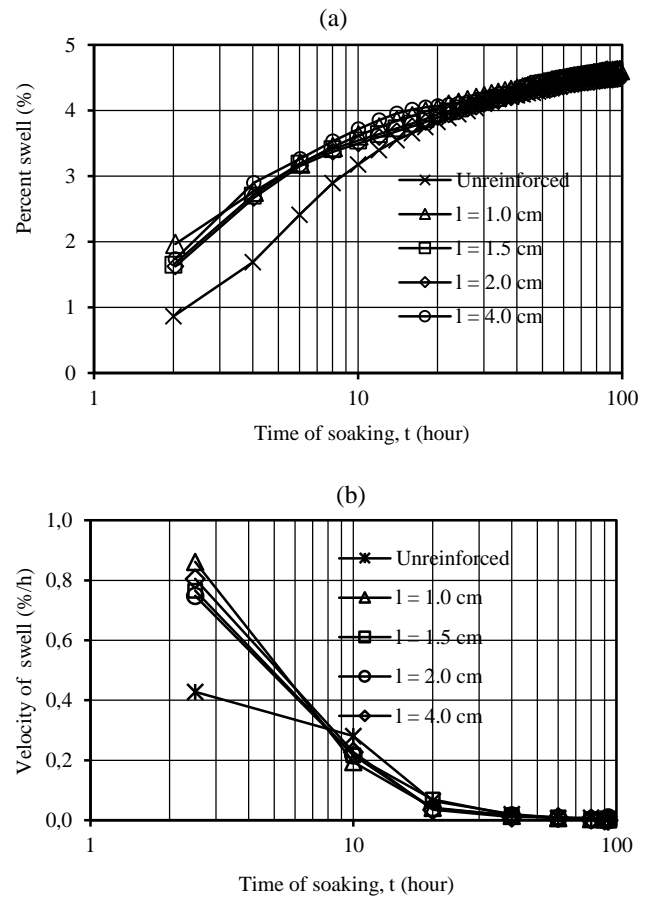


Figure 4. Swell behavior with time of unreinforced and reinforced specimens (a) percent swell and (b) velocity of swell.

TABLE II. PERCENT SWELL AND DRY UNIT WEIGHT REDUCTION AFTER 96H OF SOAKING

Thickness of sand cushion layer (mm)	Sand/Clay dry mass ratio	Final percent swell $S_{96h}$ (%)	Dry unit weight reduction $\% \Delta \gamma_d$
0	0.00	4.64	4.43
10	0.10	4.63	4.41
15	0.16	4.60	4.40
20	0.23	4.49	4.30
40	0.58	4.51	4.32

To investigate the effect of reinforcement layers on the development of swell in the reinforced specimens, the swelling velocity was evaluated as the percent swell per hour of soaking. In the first 10 hours of soaking, the reinforced specimen's swell velocity was significantly higher than that of unreinforced specimens (Fig. 4b). It could be explained by the high permeability of nonwoven geotextile layers and sand cushion, which enhancing the velocity of swell in reinforced specimens. However, after 20 hours, the influence of the reinforcement layers on the swell behavior of the reinforced specimens was diminished. The swell velocity of unreinforced and reinforced specimens reduced to less than 0.005%/h after 96h of soaking. To conclude, the geotextile- sand cushion layer induced the faster swell at the initial of soaking, but the lower final percentage of the swell.

On the other hand, during the soaking process, there are not any changes in the dry weight of soil specimens but the increment in the volume of the specimens, resulting in the decrease of dry density of the clay layers. The percentage of



dry density reduction of the clay due to 96 hours of soaking,  $\% \Delta \gamma_d$  is defined as:

$$\% \Delta \gamma_d = (\gamma_{d-\text{unsoaked}} - \gamma_{d-\text{soaked}}) / \gamma_{d-\text{unsoaked}} \times 100\% \quad (4)$$

in which  $\gamma_{d-\text{unsoaked}}$  and  $\gamma_{d-\text{soaked}}$  are the dry unit weight of clay layers before and after 96 hours of soaking, respectively

Without the consideration of thickness changes of geotextile layers and the sand cushion layer due to soaking (seem to be very small compared to that of the clay), the reduction of dry unit weight of the clay soil is evaluated using the percent swell after 96 hours of soaking,  $S_{96h}$ .

$$\% \Delta \gamma_d = 1 - 1 / (1 + S_{96h}) \quad (5)$$

As shown in Table 2, the reduction of dry unit weight of the clay in the reinforced specimens was slightly smaller than that of the unreinforced specimen. In other words, when compacted by the same density at initial, after soaking, the clay in the reinforced specimens would be higher than that in the unreinforced specimen, which contributed to the higher bearing capacity of the reinforced specimens than that of unreinforced specimens after soaking.

#### B. The CBR behavior of unreinforced and reinforced specimens

Fig 5 shows the corrected stress of the piston and penetration of un-soaked and soaked geotextile-sand cushion specimens. Compared to the unreinforced specimens, the bearing capacity of reinforced specimens was significantly higher under both soaked and unsoaked conditions. The penetrated stress increased with the increment of penetration distance. The ultimate stress in the piston was not reached within 20mm of the distance of penetration.

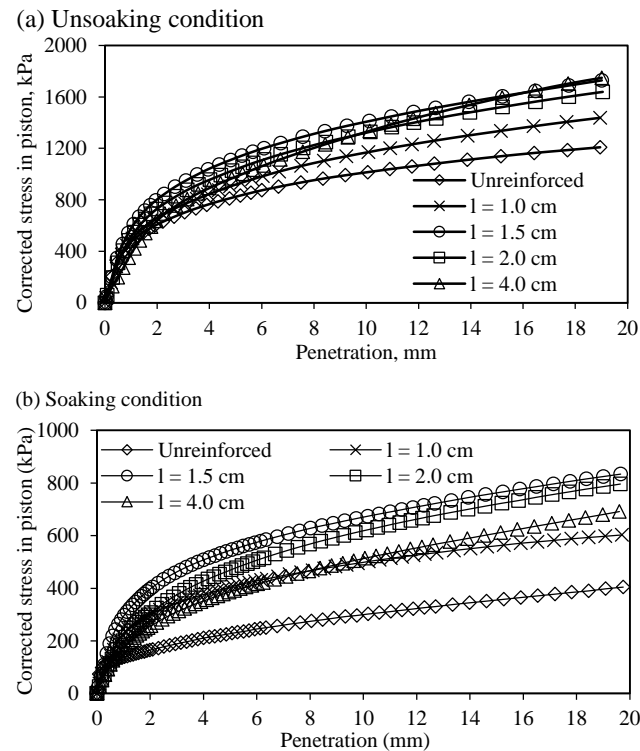


Figure 5 Corrected stress in the piston of specimen (a) without soaking and (b) soaking condition

Figure 6 showed the variation of the CBR value of specimens with the thickness of the sand cushion layer under both soak and un-soak conditions. Due to the reinforcement, the CBR value of reinforced specimens was higher than that of unreinforced specimens. Interestingly, the bearing capacity of the specimens was the highest for the specimens reinforced by 2 layers of geotextile with 1.5 cm thickness of the sand cushion, of which the ratio of the height of the topsoil layer,  $d_l$ , and the diameter of the penetrated piston,  $B$  was equal to 1. The optimum value of  $d_l/B$  was in agreement with those in previous studies. Reference [26] found that the thickness of soil required to cover geosynthetic clay liner should be at least equal to the diameter of the load piston. A similar conclusion was presented in the references [27-28] when performing the CBR test on the expansive soil subgrades reinforced with a single reinforcement layer.

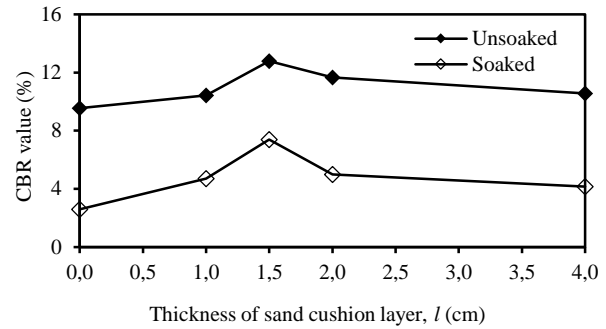


Figure 6. The variation of CBR of the soaked and unsoaked specimens with the thickness of sand cushion layer,  $l$ . The unreinforced cases are equivalent to  $l = 0$ .

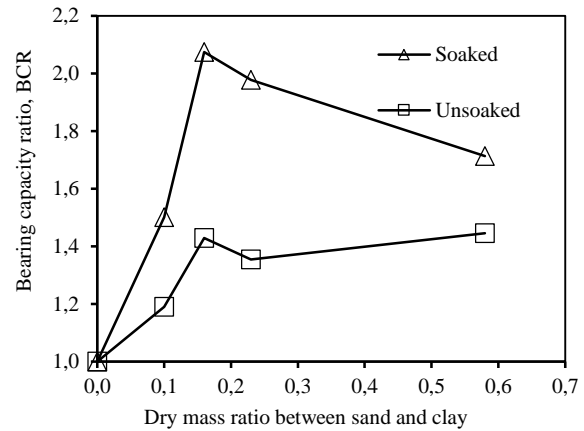


Figure 7. The correlation of strength ratio and the dry mass ratio of sand and clay

When increasing the ratio between sand and clay dry mass (Table 4), the CBR also went up in both cases (Fig. 7). For the case of un-soaking, the CBR value increased approximately 1.2 times and up to 1.4 times when the ratios were 0.1 and 0.16, respectively, compared to the un-reinforced specimen. However, the increase of the CBR value was not apparent when continuing this ratio (about 1.3 to 1.4 times when the ratio was 0.23 and 0.58). Similarly, with a larger scale for the case of the soaking process, the CBR jumped up to 1.5 and over 2 times when raising this ratio to 0.1 and 0.16 in the same order. Interestingly, for both cases, the maximum increase occurred when the ratio between sand and clay dry mass was 0.1. It can be concluded that using sand and geotextile can improve the bearing capacity of soil significantly when the soil

was wet, and the optimal dry mass ratio between sand and clay was 0.1.

### C. Influences of soaking on the CBR behavior of unreinforced and reinforced specimens

Compared to the unsoaked specimens, the CBR value of soaked specimens was much smaller, which demonstrated the extreme reduction of the strength of clay when saturated. Fig 7 shows the ratio of CBR of un-soaking and soaking specimens, which exhibited the strength reduction of specimens due to soaking. For the unreinforced specimens, the ratio reached the highest (about 3.7) and decreased to less than 2.6 for the reinforced specimens. The lowest strength reduction was 1.73 for the specimen reinforced by 1.5cm thickness of the sand cushion layer. Reference [29] also had similar observations about the significant CBR reduction when performing CBR tests after soaking at two days.

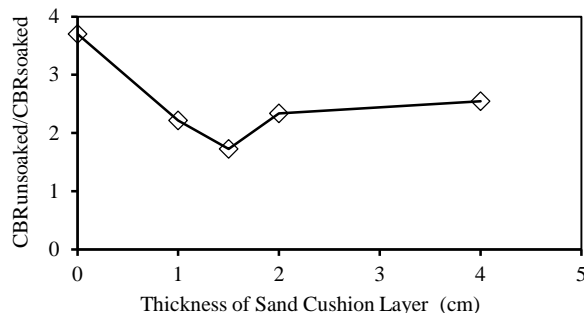


Figure 8 The influence of the thickness of sand cushion layer on the ratio of CBR of specimens before and after soaking

In short, the geotextile layer and sand cushion not only enhanced the bearing capacity of clay soil under both soaked and unsoaked conditions and minimized the strength reduction of the clayey soil after soaking.

## V. CONCLUSION

A series of CBR tests were performed to investigate the influence of geotextile and sand cushion on the bearing capacity of the soft clay. The results illustrated the critical role of the reinforcement inclusion in enhancing bearing capacity in both soaked and un-soaked conditions. The other conclusions are the following.

1) The permeable geotextile and sand cushion forced the swell to happen faster by allowing extra drainage paths into the reinforced specimens. Additionally, the density reduction fell slightly. Similarly, the percentage swell went down by over 4%.

2) It also slightly decreased the percent swell and soil density reduction after soaking.

3) The geotextile-sand cushion significantly improved the strength of soft clay for both un-soaked and soaked conditions. Based on the results of the CBR value on the 10 tested cases, the optimum thickness of sand cushion was 1cm for the 10 testing cases, which equivalent to the ratio  $d/B = 1$ .

4) When increasing the dry mass sand, the CBR value soared, particularly in the case of the soaking process. Moreover, the optimal dry mass ratio between sand and soil was 0.1 for the highest bearing capacity of the reinforced specimen under both soaked and unsoaked conditions

5) After soaking, the bearing capacity of the clay decreased significantly to 3.7 times for unreinforced specimen, while that

of reinforced specimens was less than 2.6 times, depending on the sand thickness.

Last, the significant drop of the bearing capacity when being saturated suggests that a proper function drainage system is crucial for the unreinforced and reinforced clay to maintain its bearing capacity and stabilization. For further research, the pore pressure could be measured for more detail about the soil behavior under the CBR test.

## REFERENCES

- [1] Huerta, A., & Rodriguez, A., "Numerical analysis of non-linear large-strain consolidation and filling". *Comput. Struct.* 44 (1), 357–365, 1992.
- [2] Liu, Z.Q., Zhou, & C.Y., "One-dimensional non-linear large deformation con- solidation analysis of soft clay foundation by FDM", *Acta Sci. Nat. Univ. Sunyatseni* 44 (3), 25–28, in Chinese, 2005.
- [3] Sridharan, A., Murthy, S., Bindumadhava, B.R., & Revansiddappa, K., "Technique for using fine-grained soil in reinforced earth", *Journal of Geotechnical Engineering, ASCE*, 117(8),1991, pp 1174–1190.
- [4] Chen, J., & Yu, S., "Centrifugal and numerical modeling of a reinforced lime-stabilized soil embankment on soft clay with wick drains", *International Journal of Geomechanics*, 11(3), 2011, pp 167–173.
- [5] Taechakumthorn, C. & Rowe, R., "Performance of reinforced embankments on rate-sensitive soils under working conditions considering effect of reinforcement viscosity", *International Journal of Geomechanics*, 12(4), 2012, pp 381–390
- [6] Yang, K.H., Yalaw, W.M., and Nguyen, M.D., "Behavior of Geotextile- Reinforced Clay with a Coarse Material Sandwich Technique under Unconsolidated- Undrained Triaxial Compression", *International Journal of Geomechanics, ASCE*, 16(3), 2015
- [7] Zornberg, J.G., & Mitchell, J.K., "Reinforced soil structures with poorly draining backfills. Part I: Reinforcement interactions and functions", *Geosynthetics International*, 1(2), 1994, pp 103–148.
- [8] Zhou, H., & Wen, X., "Model studies on geogrid- or geocell-reinforced sand cushion on soft soil", *Geotextiles and Geomembranes*, 26(3), , 2008, pp 231–238
- [9] Sitharam, T.G., Hegde, A., "Design and construction of geocell foundation to support the embankment on settled red mud", *Geotextiles and Geomembranes*, 41 (2013), 2013, pp 55–63.
- [10] Yu, Y., Zhang, B., & Zhang, J. M., "Action mechanism of geotextile-reinforced cushion under breakwater on soft ground". *Ocean Engineering*, 32(14-15), 2005, pp 1679–1708
- [11] Dash, S. K., & Bora, M. C., "Improved performance of soft clay foundations using stone columns and geocell-sand mattress", *Geotextiles and Geomembranes*, 41, 2013, pp 26–35.
- [12] Abdi, M. R., Sadmejad, A., & Arjomand, M. A., "Strength enhancement of clay by encapsulating geogrids in thin layers of sand". *Geotext. Geomem.*, 27(6), 2009, pp 447–455.
- [13] Abdi, M. R., & Arjomand, M. A., "Pullout tests conducted on clay reinforced with geogrid encapsulated in thin layers of sand". *Geotextiles and Geomembranes*, 29(6), 2011, pp 588–595.
- [14] Abdi, M. R., & Zandieh, A. R., "Experimental and numerical analysis of large scale pull out tests conducted on clays reinforced with geogrids encapsulated with coarse material". *Geotext. Geomem.*, 42(5), 2014, pp 494–504.
- [15] Unnikrishnan, N., Rajagopal, K., & Krishnaswamy, N.R., "Behaviour of reinforced clay under monotonic and cyclic loading", *Geotextiles and Geomembranes*, 20(2), 2002, pp 117–133.
- [16] Raisinghani, D. V., & Viswanadham, B.V.S., "Evaluation of permeability characteristics of a geosynthetic-reinforced soil through laboratory tests", *Geotext. Geomem.*, 28(6), 2010, pp 579–588.
- [17] Lin, C.Y., & Yang, K.H., "Experimental study on measures for improving the drainage efficiency of low-permeability and low-plasticity silt with nonwoven geotextile drains", *J. Chin. Inst. Civ. Hydraul. Eng.*, 26(2), 2014, pp 71–82 (in Chinese).
- [18] Minh D.N., Thanh T.N., Huu T.L., "The Effects of Soaking Process on the Bearing Capacity of Soft Clay Reinforced by Nonwoven Geotextile". In: Duc Long P., Dung N. (eds) *Geotechnics for Sustainable Infrastructure Development*. Lecture Notes in Civil Engineering, vol 62. Springer, Singapore. [https://doi.org/10.1007/978-981-15-2184-3\\_87](https://doi.org/10.1007/978-981-15-2184-3_87), 2019

- [19] Choudhary, Anil & Jha, J. & Gill, Kulbir., "A study on CBR behavior of waste plastic strip reinforced soil", *Emirates Journal for Engineering Research*. 15, 2010, pp 51-57.
- [20] Unnam Rajesh, Satish Sajja, V.K.Chakravarthi, "Studies on engineering performance of geogrid reinforced soft subgrade", *Transportation Research Procedia*, Volume 17, 2016, pp 164-173.
- [21] CA Adams, YA Tuffour, S Kwofie, "Effects of Soil Properties and Geogrid Placement on CBR Enhancement of Lateritic Soil for Road Pavement Layers", *American Journal of Civil Engineering and Architecture*, Vol. 4, No. 2, 2016, pp 62-66.
- [22] ASTM D422-63., "Standard Test Method for Particle-Size Analysis of Soils", ASTM International, West Conshohocken, PA, USA.
- [23] ASTM D698-12e2., "Standard Test Methods for Laboratory Compaction Characteristics of Soil Using Standard Effort", ASTM International, West Conshohocken, PA, USA.
- [24] ASTM D1883., "Standard Test Method for California Bearing Ratio (CBR) of Laboratory-Compacted Soils", ASTM International, West Conshohocken, PA, USA
- [25] Niteen Keerthi, Sharanabasappa Kori, "Study on Improvement of Sub Grade Soil using Soil-Reinforcement Technique", *International Journal of Applied Engineering Research*, Vol. 13, Issue 7, 2018, pp 126-134.
- [26] Koerner, R. M., & Narejo, D., "Bearing Capacity of Hydrated Geosynthetic Clay Liners", *Journal of Geotechnical and Geoenvironmental Engineering*, 121(1), 1995, pp 82-85.
- [27] Choudhary, A., Gill, K., Jha, J., & Shukla, S. K., "Improvement in CBR of the expansive soil subgrades with a single reinforcement layer", *Proc. of Indian Geotechnical Conference*, 2012, pp 289-292.
- [28] Keerthi, N., & Kori, S., "Study on Improvement of Sub Grade Soil using Soil-Reinforcement Technique", *International Journal of Applied Engineering Research* 13(7), 2018, pp 126-134
- [29] Robert G. Nini., "Effect of Soaking Period of Clay on Its California Bearing Ratio Value", *International Journal of Civil and Environmental Engineering*, Vol:13, No:2, 2019, pp 101-104

# A Strategy to Enhance Generator Efficiency of Sudoku-based PV Arrays Under Partial Shading Conditions

Le Viet Thinh

Department of Power System  
Hanoi University of Science and  
Technology  
Hanoi, Vietnam  
thinh.lv174242@sis.hust.edu.vn

Nguyen Duc Tuyen

Department of Power System  
Hanoi University of Science and  
Technology  
Hanoi, Vietnam  
tuyen.nguyenduc@hust.edu.vn

Vu Xuan Son Huu

Department of Power System  
Hanoi University of Science and  
Technology  
Hanoi, Vietnam  
huu.vxs173948@sis.hust.edu.vn

**Abstract**—Partial shading conditions (PSCs) is the most common phenomenon among PV arrays, causing significant power losses in PV arrays as well as reducing the lifespan of the photovoltaic (PV) arrays. Such method as Sudoku (SDK)-based reconfiguration can suggest a way to rearrange interconnections of PV modules of the PV array, which mitigates the effects of PSCs on the PV array. However, since this method cannot disperse the partial shading effectively in some cases, it is necessary to improve the SDK-based reconfiguration technique to achieve better output power. This paper proposes a biogeography-based algorithm to connect two adjacent PV arrays altogether to reconfigure the interconnection between PV arrays. The power-voltage (P-V) characteristics of PV arrays are drawn for a group of two PV arrays under random shading conditions. The effectiveness of the rearranging method is verified by an investigation conducted on a single-diode model-based PV model. Since the power loss is minimized 19.3% compared to the one of the conventional configured PV array and 0.3% compared to the power loss of the SDK-based configured PV array, the proposed method could be a promising solution in reducing the effects of PSCs on PV arrays.

**Keywords**—photovoltaic, reconfiguration, partial shading condition (PSC), dynamic reconfiguration, biogeography-based optimization (BBO)

## I. INTRODUCTION

The solar energy systems have been developing steadily as it offers great benefits, including the eco-friendly characteristic, zero noise, and non-production cost. Nevertheless, in the context of the exhaustion of conventional energy resources, solar energy is expected to be an essential energy resource, which contributes to the sustainable renewable energy system. Between 2019 and 2024, renewable power capacity is set to expand by 50%, increasing of 1200 GW. This trend is led by the solar photovoltaic (PV) accounting for almost 60% of the expected growth. Power generation from the solar PV was estimated to increase by more than 30% in 2018, to 580 TW [1]. However, operating a PV system has to deal with many challenging works, including solving the uncertainties of external effects. Partial shading conditions (PSCs), such as a moving cloud or shadows of neighboring PV modules, could cause PV modules of a PV array operating in heterogeneous conditions. PV modules operating under these conditions are likely to decrease the output power [2] drastically. In addition, the PV current interrupted by shaded areas could cause hotspots in PV modules reducing the lifespan of PV modules [3]. The study [3] deals with PSCs by incorporating a combination of bypass diodes and shunt resistances in the solar cell to

transmit the PV current through the shaded areas. Although this method solves the hotspots effect, it introduces multiple maximum power point (MPP) peaks on the P-V curve of the PV array, which is considered not conducive to the maximum power point tracking (MPPT) algorithm to track the MPP. Ref. [4] suggests equipping one MPPT for each PV module to track the MPP of each module, which increases the upfront capital cost. To comprehensively solve the effect of PSCs, reconfiguring the arrangement of PV modules in a PV array approach have been carried out. In reality, a PV array is composed of a number of PV modules connected with each other. In practical applications, there are five frequently used configurations of a PV array, which are simple series (SS), series-parallel (SP), bridge link (BL), honeycomb (HC), and total-cross-tied (TCT). In these five PV array patterns, TCT evidences its effectiveness in minimizing the partial shading effects [6]-[7]. A TCT PV array is parallelly connected with rows of series-connected PV modules. The effect of the irradiance on the voltage at the MPP of parallel-connected PV modules can be neglected [7]. Also, the current flowing through rows of parallel-connected PV modules will be almost proportional to the irradiance on each PV module [7]. Ideally, the output power of a PV array under PSCs is the most optimal when the irradiance values on rows of PV modules are similar.

In order to equalize the irradiance on each row of PV modules, plenty of studies propose a way to rearrange the interconnection of the PV array. All of which can be categorized into dynamic PV array reconfiguration and static PV array reconfiguration. The dynamic PV array reconfiguration alters electrical connections between PV modules of the PV array [8]. This process is achieved by using pyranometers to measure the irradiance on each PV row. From data of irradiance on each PV module of the PV array, a reconfiguration algorithm is developed to find out the optimal PV array arrangement [9]-[12]. This signal of an optimal PV array configuration is transmitted to the switching matrix, which is responsible for changing the electrical connections between PV modules into optimal PV array configuration [10]-[11]. On the other hand, the static PV array reconfiguration does not change the electrical connections but changes the physical locations of PV modules throughout the PV array. Compared to dynamic PV array configuration, static PV array configuration is more suitable to apply to large-scale PV systems as it does not require any meters, switching matrix, and only displays the switching process one time. There are a number of static PV array reconfiguration methods, such as magic-square puzzle pattern [12],

dominant-square [13], logic-based number puzzle [14], zigzag scheme [15]. Among them, Sudoku (SDK) pattern is the best configuration to alleviate the partial shading effects [16]. Although SDK-based reconfiguration proves its advantages compared to other static methods, this method does not optimally distribute the shade across the PV array in some special shading scenarios yet. The shade creates different irradiance levels on PV modules of the PV array. Therefore, the output power of the PV array does not reach the maximum value. Especially when it comes to large PV systems, this drawback of SDK-based reconfiguration becomes worse. In fact, a large PV system comprises plenty of PV arrays to reach the desired output power. Taking the advantages of two adjacent PV arrays, in this paper, a biogeography-based optimization (BBO) algorithm is proposed for reconfiguring the interconnections between PV arrays in a PV plant. This method hybridizes between SDK-based reconfiguration and BBO, called SDK-BBO-based reconfiguration, would strike the balance between the static PV array reconfiguration and dynamic PV array reconfiguration. Since the PV array is arranged based on SDK puzzle pattern, which only changes the interconnections between PV modules of PV array one-time, therefore the switching matrix is no needed for each PV array. The dynamic connections between two PV arrays gives the algorithm the flexibility in case of shading as it offers an adaptive source of irradiance to equalize the irradiance difference between PV arrays. This dynamic switching of SDK-based configured PV arrays requires less switches and switching times than the TCT reconfigured PV arrays does. Simulation results show that the SDK-BBO based reconfiguration can be applied in practical applications since it decreases the power losses from 35.8% to 16.5% compared to the one of the conventional configured PV array.

## II. SYSTEM DESCRIPTION

### A. Model a PV module

In order to evaluate the behavior of a PV cell effectively, it is needed to adopt a model representing the electrical characteristics of a PV cell. Fig. 2 displays the equivalent circuit of the single-diode model (SDM), which is frequently utilized in the field study of PV [17], [18].

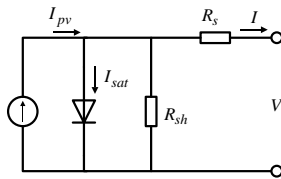


Fig. 1. Equivalent circuit of the single-diode model.

The behaviour of the solar cell is expressed by (1), which describes the I-V relationship of a PV cell.

$$I = I_{pv} - I_{sat} \left( e^{\frac{V}{aV_t}} - 1 \right) - \frac{V + IR_s}{R_{sh}} \quad (1)$$

, where  $I$  [A] and  $V$  [V] are the output current and output voltage of the PV cell, respectively.  $I_{pv}$  [A] is the PV current,  $I_{sat}$  [A] is the reverse saturation current of the diode,  $a$  [-] is the diode ideality factor,  $R_s$  [Ω] and  $R_{sh}$  [Ω] are the series and shunt resistors, respectively.  $V_t$  [V] is the thermal voltage defined by (2).

$$V_t = \frac{kT}{q} \quad (2)$$

, where  $k$  is the Boltzmann constant ( $k = 1.381 \times 10^{-23}$  [J/K]),  $T$  is the cell temperature [K],  $q$  is the electron charge ( $q = 1.60217646 \times 10^{-19}$  [C]). Five parameters of the SDM  $\{I_{pv}, I_{sat}, a, R_s, R_{sh}\}$  are obtained from the manufactured datasheet, including the values of significant points of the I-V curve, the short-circuit point, the open-circuit point, and the MPP. Solving (1) is the cardinal task to obtain the I-V characteristic of a PV cell. Herein this task is conducted in a more generic context, that is in a PV module as represented as follows.

For a PV module, (1) is modified by adding the term  $N_s$  in the exponential term, which stands for the number of PV cells in a PV module [19], [20].

$$I = I_{pv} - I_{sat} \left( e^{\frac{V}{aN_s V_t}} - 1 \right) - \frac{V + IR_s}{R_{sh}} \quad (3)$$

Substituting values of three significant points, the short-circuit point, the open-circuit point, the MPP, into (3) produces the following equations:

- At the short-circuit point ( $I = I_{sc}$ ,  $V = 0$ ) :

$$I_{sc} = I_{pv} - I_{sat} \left[ \exp \left( \frac{R_s I_{sc}}{aN_s V_t} \right) - 1 \right] - \frac{I_{sc} R_s}{R_{sh}} \quad (4)$$

- At the open-circuit point ( $I = 0$ ,  $V = V_{oc}$ ) :

$$0 = I_{pv} - I_{sat} \left[ \exp \left( \frac{V_{oc}}{aN_s V_t} \right) - 1 \right] - \frac{V_{oc}}{R_{sh}} \quad (5)$$

- At the MPP ( $I = I_{mpp}$ ,  $V = V_{mpp}$ ) :

$$I_{mpp} = I_{pv} - I_{sat} \left[ \exp \left( \frac{V_{mpp} + R_s I_{mpp}}{aN_s V_t} \right) - 1 \right] - \frac{V_{mpp} + R_s I_{mpp}}{R_{sh}} \quad (6)$$

These three equations are implicit, with five variables ( $I_{pv}$ ,  $I_{sat}$ ,  $a$ ,  $R_s$ ,  $R_{sh}$ ). To deal with this problem, an iterative method is employed as described as follows [21], [22].

The second term in (3) is assumed can be neglected [18], the photovoltaic current is rewritten as in (7).

$$I_{pv} = I_{sc} \frac{R_s + R_{sh}}{R_{sh}} \quad (7)$$

The parallel resistance is calculated by substituting (7) into (6).

$$R_{sh} = \frac{I_{sc} R_s - V_{mpp} - I_{mpp} R_s}{I_{mpp} + I_{sat} \left[ \exp \left( \frac{V_{mpp} + I_{mpp} R_s}{aN_s V_t} \right) - 1 \right] - I_{sc}} \quad (8)$$

By assuming the denominator of the right side of (8) is zero, the maximum value of the series resistance is calculated as in (9).



$$R_{s,max} = \frac{aN_s V_t \ln \left( \frac{I_{sc} - I_{mpp}}{I_{sat}} - 1 \right) - V_{mpp}}{I_{mpp}} \quad (9)$$

After that, the PV current, the reverse saturation current and the diode ideality factor are calculated by (4)-(6), respectively. This process continues with the value of the series resistance ranging from  $[0; R_{a,max}]$ . For two points in the P-V curve of a PV panel, one is on the left side, and another is on the right side, the value of the slope of the P-V curve respect to the point at the left side is greater than zero and for the point at the right side would be smaller than zero. As can be seen in Fig. 3, when the series resistance increases, the peak of the P-V curve will shift from the right side of this fixed point to the left side. Consequently, the derivative of the power with respect to the voltage will monotonically decrease from the positive to negative.

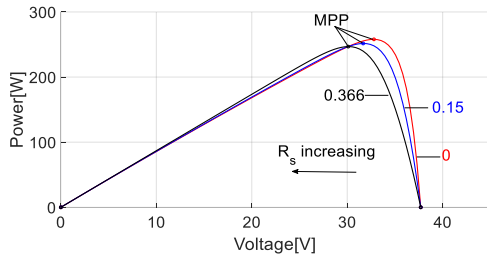


Fig. 2. P-V curves with different values of  $R_s$

As a result, the process stops until the condition (10) is satisfied.

$$\left( \frac{dP}{dV} \right)_i \left( \frac{dP}{dV} \right)_{i-1} < 0 \quad (10)$$

Using the interpolation method, the value of  $R_s$  is calculated as (11).

$$R_s = R_{s,i} + (R_{s,i} - R_{s,i-1}) \frac{\left( \frac{dP}{dV} \right)_{i-1}}{\left( \frac{dP}{dV} \right)_{i-1} - \left( \frac{dP}{dV} \right)_i} \quad (11)$$

, where  $i$  is the number of points, in which derivative of the power with respect to the voltage is greater than zero.

### B. TCT configured PV array

The TCT PV array composes a series of rows of PV modules, which are formed by parallel-connected PV modules. In this paper, the PV array chosen for study is of size  $9 \times 9$ , which is depicted as in Fig. 3. The PV array consists of 81 PV modules with nine rows and nine columns. The PV modules are labelled as ' $pq$ ' where  $p$  denotes the row and  $q$  refers to the column. For example, the module numbered 91 represents the module situated in the ninth row and first column. The current generated by a module at a specific irradiance ( $G$ -[W/m<sup>2</sup>]) is given by (12).

$$I = kI_m \quad (12)$$

, where  $I_m$ -[A] is the current generated by one PV module at standard test condition (STC) with  $G_0 = 1000$  W/m<sup>2</sup> and  $k = G/G_0$ . (12) shows the direct effect of irradiance on the input current of a PV module. Applying the KVL, the voltage of

the  $9 \times 9$  PV array is calculated by the sum of the voltages of the nine rows.

$$V_a = \sum_{p=1}^9 V_{mp} \quad (13)$$

, where  $V_a$ -[V] is the voltage of the PV array and  $V_{mp}$ -[V] denotes the voltage of the PV modules at the  $p^{th}$  row.

For a row of PV modules, the sum of the current is the summation of current limited on each individual PV module. Hence the current of any row in  $9 \times 9$  PV array is calculated as:

$$I_{Rn} = \sum_{n=1}^9 kI_{1n} \quad (14)$$

The current at each node in the array can be calculated by applying Kirchhoff's Current Law:

$$I_a = \sum_{q=1}^9 (I_{pq} - I_{(p+1)q}) = 0, p = 1, 2, 3, \dots, 8. \quad (15)$$

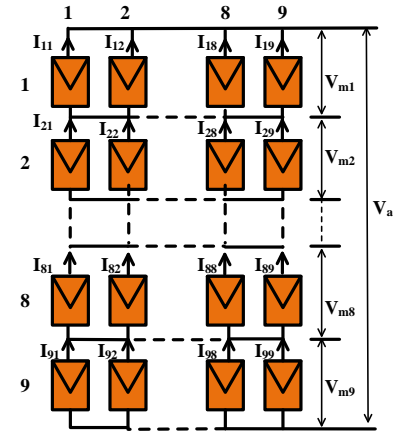


Fig. 3.  $9 \times 9$  TCT configured PV array

### C. Sudoku-based configured PV array

Sudoku is a logic-based number-placement puzzle. Solving a Sudoku puzzle involves logically placing numbers from 1 to 9 into a  $9 \times 9$  matrix, which is divided into 9 ( $3 \times 3$ ) sub-matrices. Each  $3 \times 3$  sub-matrix has to be accommodated the digits 1 to 9 so that each row, column and sub-matrix  $3 \times 3$  contains the numbers 1 to 9 without repeating any numbers. Inspired by the rule of Sudoku puzzle, a procedure for rearranging the configuration of PV array under PSCs is established to disperse the shade across. This procedure is done with the aim to minimize the difference between irradiance on rows of PV modules of the PV array. The logic sequence is to develop the puzzle for a  $9 \times 9$  PV array as follows:

- The first column of the Sudoku is filled by numbers 1 to 9 in succession.
- To avoid the repetition of numbers of the previous column, the second and third columns are established by shifting the numbers in the previous column by three.
- If the fourth column is shifted again by three, the arrangement of this column will be the same as in the first column. This is not the intention of the

arrangement as any two elements of the same row must be unique. Hence the numbers in the previous column are shifted by three, but the middle digit is replaced by the first digit and vice versa.

- The fifth and the sixth column are filled by shifting the numbers in the previous column by three.
- For the seventh column, the numbers in the previous column are shifted by three, and then the middle digit is replaced by the first digit and vice versa.
- The last two columns are filled by shifting the previous column by three.

### III. BIOGEOGRAPHY-BASED OPTIMIZATION IMPLEMENTATION TO PV ARRAY RECONFIGURATION

#### A. Biogeography-based optimization

The BBO algorithm is an evolutionary algorithm proposed by D. Simon [23]. A BBO model describes the way of species migrate from one habitat to another, evolve and diminish the population. A geographical area is evaluated by habitat suitability index (HSI), which presents how well suited for biological species to live in. HSI correlates with features such as rainfall, diversity of vegetation, land area, diversity of topographic features, and temperature.

Habitats with a high HSI tend to have a large number of species and have a low species immigration rate as they are already almost saturated with species. For the same reason, species in high HSI habitats have a high emigration rate because the current habitats are so competitive.

On the other hand, habitats with low HSI have a small number of species. These habitats have more opportunities for species from other locations to land here, thus the immigration rates of these habitats are high. The immigration rates of low HSI habitats are low as species tend to stay at their home rather than move to other habitats.

Biogeography is nature's way of distributing species, which is implemented to obtain the optimal configuration of the PV array. The evaluation progress of BBO is made by mimicking the migration and mutation process in biogeography.

#### B. Standard Deviation

Like any evolutionary algorithm, the accuracy of the BBO algorithm largely depends on the definition of the fitness function. Therefore, to increase the probability of finding the optimal configuration, the fitness function needs to be defined properly. Herein, the standard deviation is utilized as a fitness function. The standard deviation is a measure of the variation or the dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range. By minimizing the SD of the rows currents, the mismatching effect between PV modules is reduced, thus eliminating multiple peaks on the P-V curve of the PV array and increasing the output power of the PV array. Herein, the objective function for the chosen interconnections between two PV arrays can be defined as follows.

$$\sigma = \sqrt{\frac{1}{9} \left[ \sum_{j=1}^9 (I_j - I_m)^2 \right]} \quad (16)$$

, where  $I_j$ -[A] is the current of the  $j^{\text{th}}$  column,  $I_m$ -[A] is the current of one column at the STC ( $G = 1000 \text{ W/m}^2$ ).

#### C. BBO implementation to PV array reconfiguration

The following pseudo code represents how BBO is implemented to the PV array reconfiguration.

##### **Initialize input parameters:**

*Initialize  $N$  irradiance patterns  $\{x_k\}$ ; population size; generations; crossover and mutation operator, elite rate.*

##### **While not** (termination criterion)

*Generate initial population  $x_k$*

*For each  $x_k$ , set emigration probability ( $\alpha_k$ ), immigration probability ( $\beta_k = 1 - \alpha_k$ )*

*$\{z_k\} \leftarrow \{x_k\}$*

*For each  $z_k$  ( $k=1, \dots, N$ ) do*

*For each independent variable index  $s \in [1, n]$  do*

*Use  $\beta_k$  to probabilistically decide whether to migrate to  $z_k$*

*If immigrating then*

*Use  $\alpha_k$  to probabilistically select the emigrating individual  $x_j$*

*$z_k(s) \leftarrow x_j(s)$*

*End if*

*Next independent variable index:  $s \leftarrow s + 1$*

*Probabilistically mutate  $z_k$*

*Next individual:  $k \leftarrow k + 1$*

*$\{x_k\} \leftarrow \{z_k\}$*

*Next generation*

### IV. RESULTS AND DISCUSSION

To evaluate the BBO performance on reconfiguring connections of PV arrays, two case studies are conducted on two adjacent PV arrays, simulated with the random partial shades. In case study 1, the first PV array is under the short-narrow shade, and the other is under short-wide shade. Case study 2 is formed by the first PV array is under the long-wide shade, and the other is under long-narrow shade. The shades have different levels of irradiance, from  $200 \text{ W/m}^2$  to  $1000 \text{ W/m}^2$ . Fig. 4 depicts the experimental PV arrays under the short narrow-short wide (SN-SW) shade. Fig. 4a represents the allocation of the shade on TCT configured PV arrays. After applying the SDK-based reconfiguration technique, the shade is dispersed as in Fig. 4b. Although the shade is distributed across PV arrays, there are differences between the irradiance of rows of PV modules. The connections between two PV arrays are rearranged to as in Fig. 4c. Analogously, Fig. 5 expresses PV arrays under the long wide-long narrow (LW-LN) shade with three configurations.

Fig. 6 demonstrates P-V curves of two PV arrays under the uniform condition and two shading scenarios, with three configurations, TCT configuration, SDK-based

configuration, and SDK-BBO-based configuration. When there is no shading on PV arrays, the irradiance on PV arrays is uniform, thus the P-V curve appears only one MPP. Under PSCs, the P-V curve of the conventional TCT configured PV arrays has multiple MPPs. Besides the global MPP (GMPP), the P-V curves also have one to two local MPPs. It is observed that the GMPP of TCT configured PV arrays is much lower than the GMPP of PV arrays under uniform condition, which shows the detrimental effects of PSCs on the PV array. However, after rearranging by the SDK-based reconfiguration and SDK-BBO-based reconfiguration, the P-V curve becomes more “smooth”. The GMPP of the P-V curve of SDK-BBO reconfiguration is higher than the one applied only SDK-based reconfiguration.

In respect of quantitative assessment, the performance of the PV array reconfiguration method under PSCs is evaluated by comparing the values of the GMPP, power loss ( $PL$ ), fill factor ( $FF$ ), and efficiency ( $E$ ). The mismatching

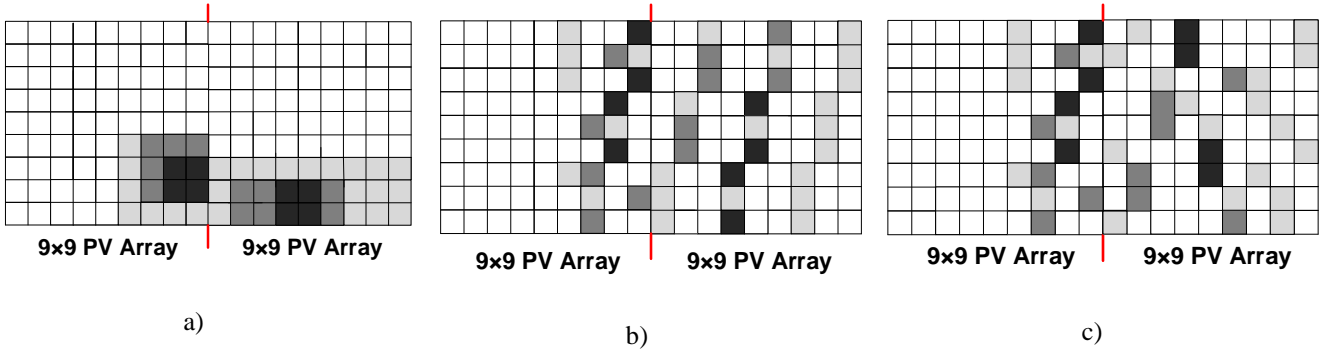
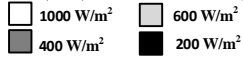


Fig. 4. Simulation PV arrays under SN-SW shade: a) TCT PV arrays; b) SDK-based configured PV arrays; c) SDK-BBO-based configured PV arrays

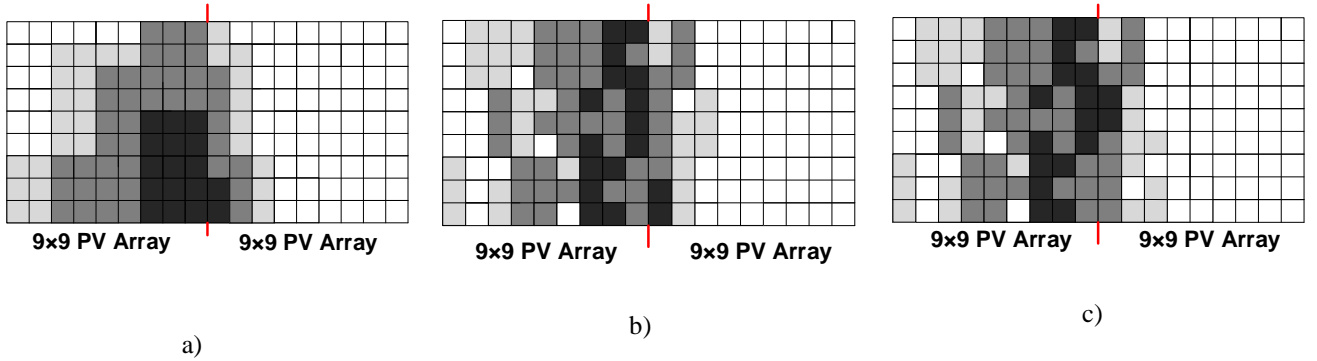


Fig. 5 Simulation PV arrays under LW-LN shade: a) TCT PV arrays; b) SDK-based configured PV arrays; c) SDK-BBO-based configured PV arrays

power loss ( $\Delta P_L$ [%]) of the PV array is found by (17) [24].

$$\Delta P_L = \frac{P_U - P_{PSC}}{P_U} \times 100 \quad (17)$$

, where  $P_U$ -[W] is the maximum power of PV array under uniform condition,  $P_{PSC}$ -[W] is the maximum power under PSCs. The fill factor is given as

$$FF = \frac{P_{max}}{V_{oc} \times I_{sc}} \quad (18)$$

, where  $P_{max}$ -[W] is the maximum power of the PV array under a given condition. Efficiency ( $E$ [%]) is defined as the ratio of the maximum power to the input power, which is calculated as the incoming irradiance on the total square of PV array as in (19).

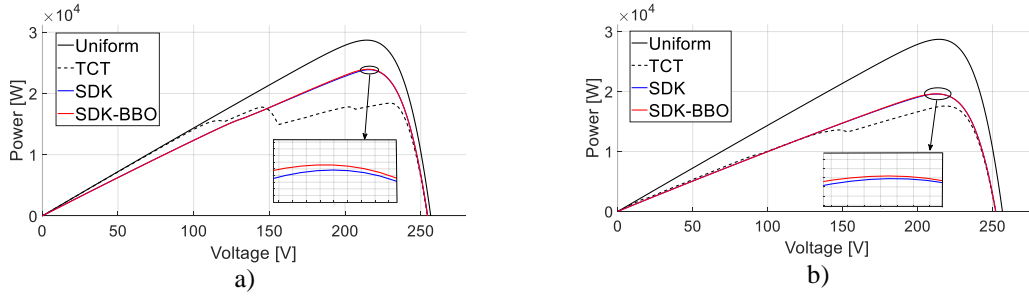


Fig. 6. P-V curve of PV arrays under uniform irradiance and PSCs: a) SN-SW; b) LW-LN

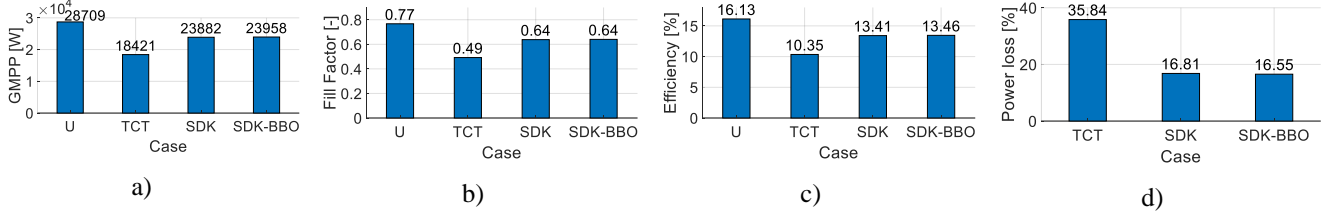


Fig. 7. Performance evaluation of PV arrays under SN-SW shade: a) GMPP; b) FF; c) E; d) PL

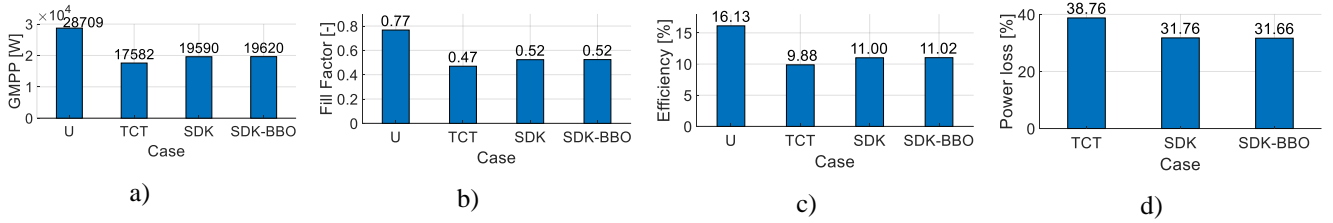


Fig. 8. Performance evaluation of PV arrays under LW-LN shade: a) GMPP; b) FF; c) E; d) PL

$$E = \frac{P_{\max}}{S_{PV \text{ array}} \sum_{i=1}^{117} G_i} \times 100 \quad (19)$$

, where  $G_i$  [W/m<sup>2</sup>] is the irradiance on each PV module of the PV array,  $S_{PV \text{ array}}$  [m<sup>2</sup>] is the total square of the PV array.

As can be seen from Fig. 7, the MPP of two SDK-BBO configured PV arrays increases 5537 W compared to the TCT configured PV arrays and 76 W compared to two SDK configured PV arrays. The mismatching power loss of SDK-BBO configured PV arrays decreases from 35.84% to 16.55% compared to TCT configured PV arrays. The SDK-BBO arrangement is more effective than the TCT PV array as it enhances the FF of the PV array from 0.49 to 0.64. The efficiency of SDK-BBO configured PV arrays increases from 13.41% to 13.46% compared to the SDK configured PV arrays.

Similarly, in Fig. 8, all the performance metrics of two SDK-BBO configured arrays are better than the ones TCT configured PV arrays and SDK configured arrays. The GMPP of the SDK-BBO configured PV arrays increases from 19590 W to 19620 W compared to the SDK configured PV arrays.

Shortly, the data in Fig. 7 and Fig. 8 are evident that four criteria (GMPP, PL, FF, E) are optimized after suggesting the SDK-BBO-based arrangement. Although the performance indices of SDK-BBO based reconfiguration are not so large compared to the ones of the SDK-based reconfiguration in the chosen tests, the improvement can magnify when it is applied to a large PV system consisting of thousands of PV arrays.

Further research should be conducted with real data of irradiance of PV arrays under PSCs since the variation of irradiance, in reality, is sharper than the one in this research. The variation of irradiance would take the advantages of this method as the adaptive source of irradiance is more flexible.

## V. CONCLUSIONS

In this paper, a BBO-SDK reconfiguration is proposed to arrange the connections between PV arrays under PSCs. A simulation of two adjacent PV arrays under PSCs is established. This model is constructed based on the SDM to validate the performance of the proposed reconfiguration method. The effectiveness of the method is proven by comparing the power loss, fill factor and efficiency of the proposed reconfiguration to the ones of the SDK reconfiguration and the TCT configuration. Since the output power is optimized compared to the TCT PV array, this arrangement could be applied to reconfiguring the connections between PV arrays of a PV system, especially the large one.

## ACKNOWLEDGMENT

This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2020-SAHEP-005.

## REFERENCES

- [1] G. Walker, "Evaluating MPPT converter topologies using a matlab PV model," J. Electr. Electron. Eng. Aust., vol. 21, no. 1, pp. 49–55, 2001.
- [2] J. A. Gow and C. D. Manning, "Development of a photovoltaic array model for use in power-electronics simulation studies," IEE Proc. Electr. Power Appl., vol. 146, no. 2, pp. 193–200, 1999.

- [3] C. Science, "Effect of shunt resistance and bypass diodes on the shadow tolerance of solar cell modules," vol. 5, pp. 183–198, 1982.
- [4] S. Malathy and R. Ramaprabha, "Performance enhancement of partially shaded solar photovoltaic array using grouping technique," *J. Sol. Energy Eng. Trans. ASME*, vol. 137, no. 3, pp. 1–5, 2015.
- [5] S. R. Pendem and S. Mikkili, "Modelling and performance assessment of PV array topologies under partial shading conditions to mitigate the mismatching power losses," *Sol. Energy*, vol. 160, no. November 2017, pp. 303–321, 2018.
- [6] F. Belhachat and C. Larbes, "Modeling, analysis and comparison of solar photovoltaic array configurations under partial shading conditions," *Sol. Energy*, vol. 120, pp. 399–418, 2015.
- [7] G. Velasco-Quesada, F. Guinjoan-Gispert, R. Piqué-López, M. Román-Lumbreras, and A. Conesa-Roca, "Electrical PV array reconfiguration strategy for energy extraction improvement in grid-connected PV systems," *IEEE Trans. Ind. Electron.*, vol. 56, no. 11, pp. 4319–4331, 2009.
- [8] D. Nguyen and B. Lehman, "An adaptive solar photovoltaic array using model-based reconfiguration algorithm," *IEEE Trans. Ind. Electron.*, vol. 55, no. 7, pp. 2644–2654, 2008.
- [9] T. N. Ngoc et al., "A hierarchical architecture for increasing efficiency of large photovoltaic plants under non-homogeneous solar irradiation," *Sol. Energy*, vol. 188, no. January, pp. 1306–1319, 2019.
- [10] T. Ngo Ngoc, Q. N. Phung, L. N. Tung, E. Riva Sanseverino, P. Romano, and F. Viola, "Increasing efficiency of photovoltaic systems under non-homogeneous solar irradiation using improved Dynamic Programming methods," *Sol. Energy*, vol. 150, pp. 325–334, 2017.
- [11] M. Matam and V. R. Barry, "Improved performance of Dynamic Photovoltaic Array under repeating shade conditions," *Energy Convers. Manag.*, vol. 168, no. November 2017, pp. 639–650, 2018.
- [12] A. S. Yadav, R. K. Pachauri, Y. K. Chauhan, S. Choudhury, and R. Singh, "Performance enhancement of partially shaded PV array using novel shade dispersion effect on magic-square puzzle configuration," *Sol. Energy*, vol. 144, pp. 780–797, 2017.
- [13] B. Dhanalakshmi and N. Rajasekar, "Dominance square based array reconfiguration scheme for power loss reduction in solar PhotoVoltaic (PV) systems," *Energy Convers. Manag.*, vol. 156, no. September 2017, pp. 84–102, 2018.
- [14] H. S. Sahu, S. K. Nayak, and S. Mishra, "Maximizing the Power Generation of a Partially Shaded PV Array," *IEEE J. Emerg. Sel. Top. Power Electron.*, vol. 4, no. 2, pp. 626–637, 2016.
- [15] S. Vijayalekshmy, G. R. Bindu, and S. Rama Iyer, "A novel Zig-Zag scheme for power enhancement of partially shaded solar arrays," *Sol. Energy*, vol. 135, pp. 92–102, 2016.
- [16] G. Sai Krishna and T. Moger, "Reconfiguration strategies for reducing partial shading effects in photovoltaic arrays: State of the art," *Sol. Energy*, vol. 182, no. July 2018, pp. 429–452, 2019.
- [17] A. N. Celik and N. Acikgoz, "Modelling and experimental verification of the operating current of mono-crystalline photovoltaic modules using four- and five-parameter models," *Appl. Energy*, vol. 84, no. 1, pp. 1–15, 2007.
- [18] J. Cubas, S. Pindado, and M. Victoria, "On the analytical approach for modeling photovoltaic systems behavior," *J. Power Sources*, vol. 247, pp. 467–474, 2014.
- [19] A. Laudani, F. Riganti Fulginei, and A. Salvini, "Identification of the one-diode model for photovoltaic modules from datasheet values," *Sol. Energy*, vol. 108, pp. 432–446, 2014.
- [20] A. Hussein, "A simple approach to extract the unknown parameters of PV modules," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, no. 5, pp. 4431–4444, 2017.
- [21] N. D. Tuyen, N. D. Huy, L. V. Thinh, and H. T. Thanh, "An explicit approach to simulate the five- parameter model for PV panels under various conditions," *SEATUC Journal of Science & Engineering*, vol. 1, no. 2, pp. 6–13, 2020.
- [22] T. Nguyen-Duc, H. Nguyen-Duc, T. Le-Viet, and H. Takano, "Single-diode models of PV modules: A comparison of conventional approaches and proposal of a novel model," *Energies*, vol. 13, no. 6, 2020.
- [23] S. Mirjalili, "Biogeography-based optimisation," *Stud. Comput. Intell.*, vol. 780, no. 6, pp. 57–72, 2019.
- [24] A. Mäki, S. Valkealahti, and J. Leppäaho, "Operation of series-connected silicon-based photovoltaic modules under partial shading conditions," *Prog. Photovoltaics Res. Appl.*, vol. 20, no. 3, pp. 298–309, 2012.



# The Influence of Water Content and Compaction on the Unconfined Compression Strength of Cement Treated Clay

Minh-Duc Nguyen

Faculty of Civil Engineering  
Ho Chi Minh City University of Technology  
and Education  
Ho Chi Minh City, Vietnam  
ducnm@hcmute.edu.vn

Anh-Thang Le

Faculty of Civil Engineering  
Ho Chi Minh City City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
lathang@hcmute.edu.vn

Thien-An Nguyen

Faculty of Civil Engineering  
Ho Chi Minh City City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
nta1701@gmail.com

Nguyen-Thao Thach

Faculty of Civil Engineering  
Ho Chi Minh City City University of Technology and Education  
Ho Chi Minh City, Vietnam  
thaotn.cons@gmail.com

Thanh-Kiet Phan

Faculty of Civil Engineering  
Ho Chi Minh City City University of Technology and Education  
Ho Chi Minh City, Vietnam  
17149022@student.hcmute.edu.vn

**Abstract**— The paper presents a series of laboratory tests for investigating the influence of water content on the unconfined compression strength (UCS) of compacted clay treated by cement. The mixture of dry clay, 10% of cement in weight, and different water content were compacted to make the reinforced specimens. The specimens were then determined the UCS after 7, 14, and 28 days of curing without the changes of their wet mass. The test results revealed that compared to the compacted clay, optimum moisture content to achieve the highest density of the compacted mixture was smaller by 5%. Apart from the unreinforced specimens, the UCS of cement-treated clay specimens increased with the increment of curing days due to the hydration process of cement. Based on the test results, to obtain the highest UCS development, the water content of the mixture should be as high as 35-40%. After 28 days of curing, the reinforced specimens reached the maximum UCS when compacted and mixed by 35% of water content, which was higher than the OMC of the mixture by 5%.

**Keywords**— Compaction, Cement treated clay, water content

## I. INTRODUCTION

The soft clay treated by cement was extensively studied in many previous studies. A binding agent, cement, was mixed with soft clay to procedure a mechanical stable soil matrix to increase the strength, stiffness, reduce the water void, and immobilize possible contaminants [1-2]. During the solidification of the cement-treated clay, the water liquid of specimens reduces due to both evaporation and the formation of hydration water [3-4], which induces the strength improvement of the cement-treated clay (Figure 1). As a result, the water content of the mixture would reduce while the dry unit weight might increase, resulting in the development of its strength due to the hydration process. Like concrete materials, several studies observed the increment in unconfined compressive strength, UCS of the cement-treated clay with the days of curing [5].

Moreover, the compaction process was also applied to strengthen the cement-treated clay further. Since the structure of cement admixed clay was might destroyed under Proctor compaction [5-6], in general, the cement-clay mixture is compacted immediately after mixing clay with cement binder

to avoiding the solidification before compaction [4, 7-8]. The compaction behavior of the mixtures was significantly changed by the cement binder. The optimum moisture content, OMC of lateritic soil with 2-10% of cement binder (in weight) was smaller, while the maximum dry density of the compacted mixture was higher than that of the pure soil [9]. The increment of maximum dry density was also observed in an expansive clay stabilized with and without lime sludge and 10% cement [4].

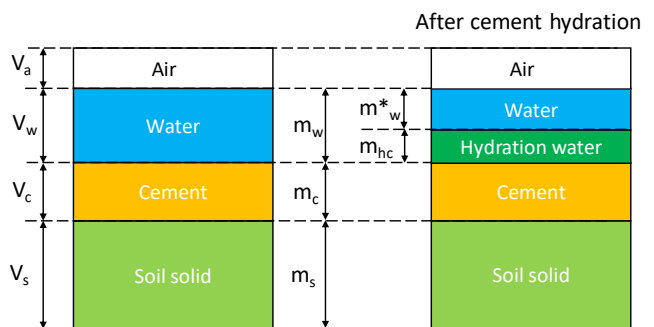


Figure 1. The changes of phase diagram in cement-treated clay due to the solidification process

The influence of water content on the strength development of the cement-treated clay has been studied previously. The water-soil transfer mechanism was analyzed using the three forms of water, including bound water, free water, and hydration water in the solidified dredged materials (DM), which are the water molecules, loosely held water, and chemically bound water, respectively [3,10]. The test results showed that the unconfined compressive strength of the solidified DM was related consistently to either hydration water or bound water, depending on the percent of cement. The influence of water content on the UCS of the soil-cement mixture was investigated in [11]. The test results revealed that there was an explicit optimum water content providing the maximum UCS value of clay treated by different percent of cement. Last, the higher water to cement ratio was adapted, the higher rate of compressive strength development.

In what it concerns compaction, the soil specimens should be compacted at optimum moisture content to reach the maximum dry density. On the other hand, the solidification process of the cement soil mixtures might be accelerated at higher water content. The water content should be kept sufficient for the optimum solidification of cement without inducing the strength reduction caused by the appearance of water between particles. In other words, the water content for cement-treated clay compaction should be chosen based on the consideration of dry density combining with the UCS and UCS improvement of specimens due to the solidification process. In most of the previous studies, the cement-treated soil was compacted at its optimum moisture content to achieve the highest MDD without considering the influence of water content on the formation of hydration water and the generation of UCS of specimens after curing [8-9]. Thus, the maximum strength improvement of cement-treated soil might not be observed if it is compacted at the optimum moisture content without adding any additional water during the curing process. The objectives of this study were to (1) investigate the UCS behavior, and (2) determine the optimum moisture content for cement-treated clay compaction to produce the highest UCS and UCS improvement of specimens after 28 curing days.

## II. TEST MATERIALS

### A. Soft clay

The soft clay was excavated from Saigon river, Ho Chi Minh City, Vietnam. It is classified as a low plastic inorganic silt (MH) by the Unified Soil Classification System (Figure 2). Its liquid limit (LL), plastic limit (PL), and specific gravity are 43.6, 31.1, and 2.73, respectively. Table 1 shows the results of the standard Proctor test, in which the value of optimum moisture content and the maximum dry unit weight are 25.2% and 14.47 kN/m<sup>3</sup>, respectively.

TABLE I. PROPERTIES OF SAIGON RIVERBED CLAY

Parameters	Value
Percent of finer content	82.74
Percent of % sand	17.26
Plastic limit, PL	31.1
Liquid limit, LL	43.6
Plasticity index, PI	12.5
Specific gravity, G <sub>s</sub>	2.73
Optimum moisture content, %	25.2
Maximum dry unit weight, kN/m <sup>3</sup>	14.47
USCS	MH

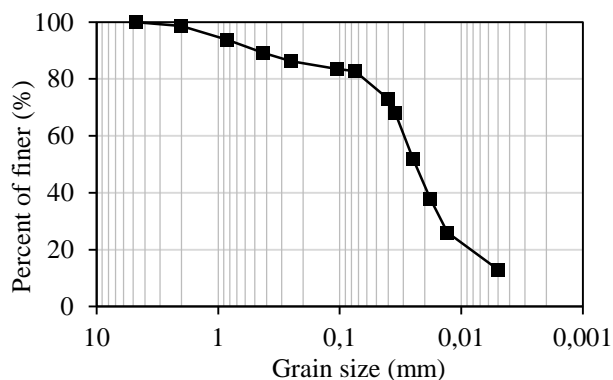


Figure 2. The grain size distribution of the Saigon riverbed clay

### B. Poland cement

Portland cement type PC40 was used in this research. The properties of cement are summarized in the following table. Table 2 shows the results of the minimum compressive strength of 40 MPa, 28 days after the date of manufacture.

TABLE II. PROPERTIES OF POLAND CEMENT

Cement properties	Value
Minimum compressive strength, MPa:	
- 3 days $\pm$ 45 min	21
- 28 days $\pm$ 8 hours	40
Setting time, minimum values	
- The initial setting time (minutes)	45
- The final setting time (minutes)	375
Fineness:	
- The retaining on the sieve size 0.09 mm (%)	10
- The Blaine surface area (cm <sup>2</sup> /g)	2800
Le-chatelier apparatus test (mm)	10

## III. EXPERIMENTAL PROGRAM

A total of 75 laboratory tests was conducted to determine the unconfined compression strength of the unreinforced and cement-treated clay. The test variation included the water content of specimens and the curing time after compaction.

### A. Specimen preparation

After being excavated from the riverbed, the soft clay was dried in an oven (less than 60°C) for about 24h then crushed and ground into powder in a mortar. For the unreinforced specimens, the dried powder clay then mixed with water corresponding to 20, 25, 30, 35, and 40%, stored in a resealable plastic bag within a temperature-controlled chamber for a minimum of a day to ensure a uniform distribution of moisture within the soil mass. The unreinforced specimens for unconfined compression strength tests were compacted in a cylindrical mold 50mm in diameter and 100mm in height. The soil was placed in 3 layers with compaction by manual tamping. The dry unit weight of the specimens was controlled to be the same as that obtained from the standard compaction test.

For the cement-treated clay specimens, the specimens were prepared by mixing the dry powder clay with cement. The content of cement was set at 10% in dry mass, which was the cement content for the high UCS performance of cement-treated soil proposed in the previous studies. An addition of 10% of cement optimally improved the California bearing ratio and the unconfined compression strength of the cement-treated lateritic soils after compaction [9]. At the cement content higher than 8%, the favourable compressive strength of cement stabilized clay sandy soil was observed at the dry state and after 48h immersion in water [8]. Before preparing the specimens for the UCS test, the dried powder mixture was mixed well with various amounts of water for the standard Proctor compaction test. The mixing was done as fast as possible to avoid the hardening of the cement-soil mixture. The obtained compaction results of the dry unit weight at different the water content of the clay – cement mixture were then applied to prepare the cement-treated soil specimens for the UCS test. The compaction process to make those

specimens was similar to that of unreinforced specimens, which were performed quickly after mixing. The cylindrical specimens were then covered by plastic wraps to keep the constant weight and cured for 7-days, 14-days and 28-days in an environmental chamber where the ambient temperature and relative humidity were maintained at  $25 \pm 1^\circ\text{C}$  and higher than 95%, respectively. After curing, the weight of specimens was measured to determine the decrement of specimen weight due to evaporation. The results showed that the lost weight of specimens still kept unchanged with less than 0.1% of weight decrement

#### B. Testing program

The laboratory test for unconfined compression strength was performed following [12] in which the load was applied to produce an axial deformation rate at a rate of 1mm/min (i.e., axial strain rate at 1%/min). The compressive stress  $\sigma_c$  was calculated using the average cross-sectional areal,  $A$ , and the applied load,  $P$ .

$$\sigma_c = \frac{P}{A} \quad (1)$$

in which the average cross-sectional area could be evaluated based on the initial cross-sectional area,  $A_0$ , and the axial strain,  $\varepsilon_1$ .

$$A = \frac{A_0}{1 - \varepsilon_1} \quad (2)$$

Prior to testing, the top and bottom specimens were trimmed to make a perfectly flat surface. At least three replicate specimens were prepared and tested for the unconfined compression strength. The mean UCS values of unreinforced and reinforced clay specimens were shown in Figure 7, of which the coefficient of variation (COV) is less than 0.1.

### IV. RESULTS AND DISCUSSION

#### A. Compaction behavior of cement – clay mixture

The compaction curves of unreinforced clay and cement – clay mixture are shown in Figure 3. The optimum moisture content (OMC) and the maximum dry density (MDD) of the mixture of clay with 10% cement are 23.7% and 13.70 kN/m<sup>3</sup>, respectively, which are smaller than those of the unreinforced clay specimens. The changes of OMC due to the addition of cement binders were reported inconsistently in the previous studies. The reduction of OMC with the increment in MDD was observed when adding cement binder into sandy soil [13]. On the other hand, the increment in the OMC and decrement in MDD of the cement soil mixture was reported in [14-15]. By contrast, the value of OMC and MDD were found to be increased for an expansive clay and 10% cement [4]. The inconsistent trending changes might be due to the difference in the soil types and cement binders, which significantly influence the compaction behavior of the cement-soil mixture.

In addition, the degree of saturation of the compacted specimens was also evaluated, as shown in Figure 4. When increasing water content, the degree of saturation of compacted specimens increased and reached over 93%, then remained constant. A similar degree of saturation behavior was also found in the compaction clay with and without geosynthetic reinforcement layers [16].

The optimum degree of saturation of compacted specimens  $S_{r(opt)}$  is defined as  $S_r$  at which the MDD is obtained [17]. The  $S_{r(opt)}$  value of unreinforced clay is 80.4%, which is close to the average  $S_{r(opt)}$  value of 82% proposed by [17]. While the  $S_{r(opt)}$  value of clay significantly dropped to 66.7% when mixing with 10% cement.

#### B. Unconfined compressive strength behavior of cement-treated clay

The stress-strain behaviors of the unreinforced clay and cement-treated clay specimens at different water content are shown in Figures 5 and 6, respectively. For the unreinforced specimens, the higher the water content, the more ductile stress-strain behavior was found. In other words, the axial strain at failure would be higher for the specimens at the higher water content. It is interesting to observe that the unreinforced clay specimens exhibited the highest UCS at 180 kPa when prepared at OMC-5%. The UCS of specimens was then significantly dropped to less than 31 kPa when increasing the water content to be higher than OMC+10%.

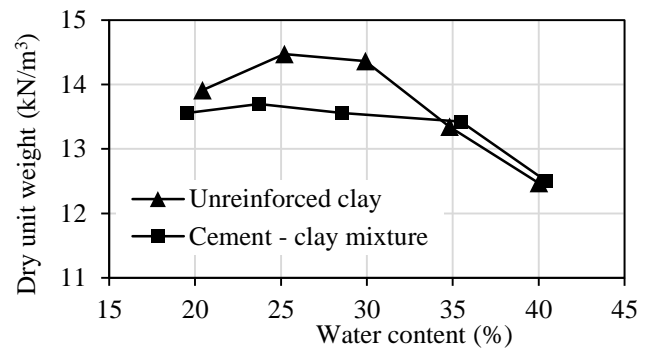


Figure 3. Compaction curves of unreinforced clay and cement clay mixture

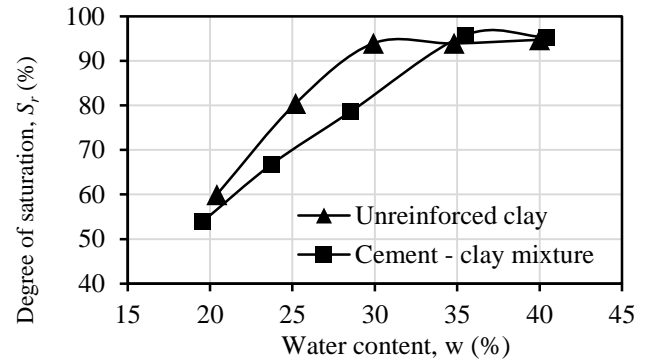


Figure 4. The degree of saturation of unreinforced and cement – clay mixture after compaction

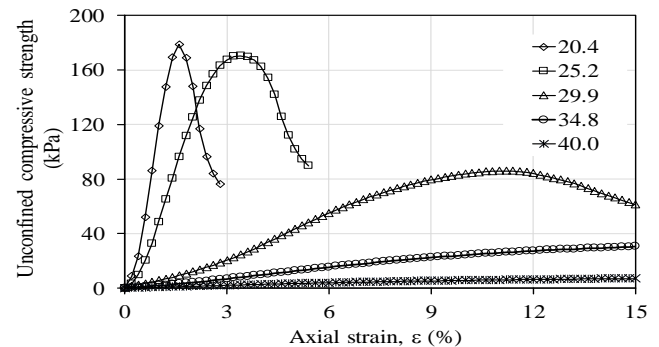


Figure 5. The UCS behavior of unreinforced clay at different water contents

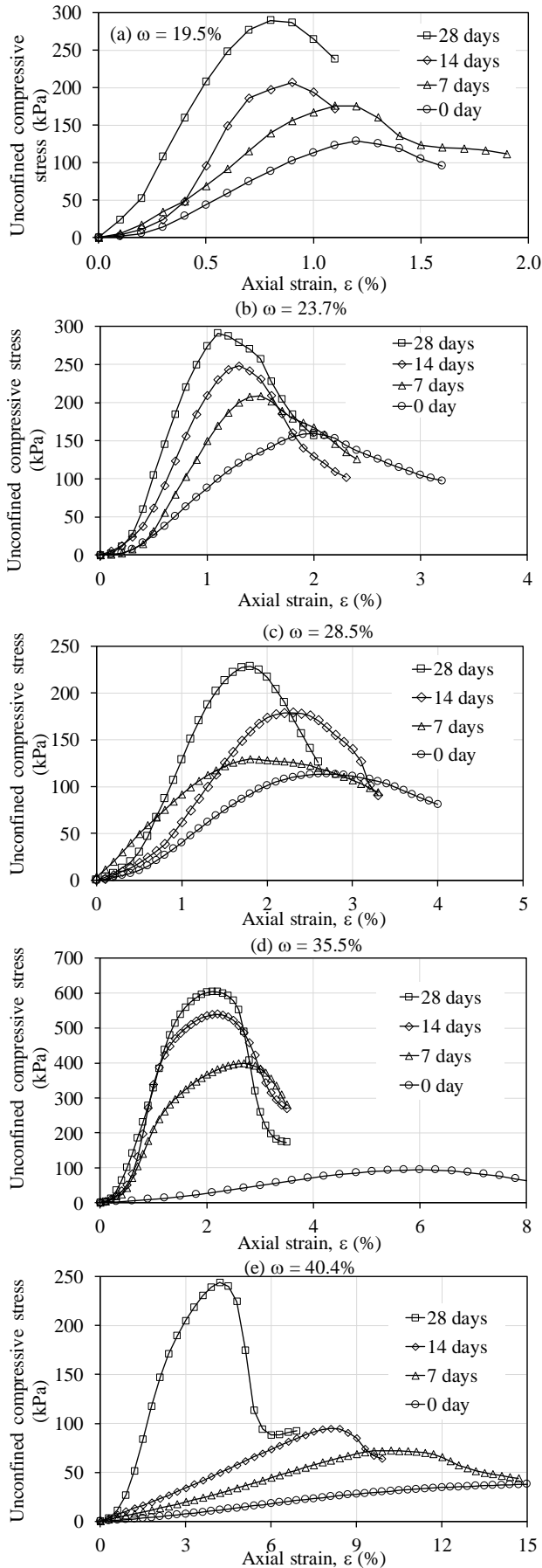


Figure 6. The UCS of cement-treated clay at different initial water content (a) 19.5%; (b) 23.7%; (c) 28.5%; (d) 35.5% and (e) 40.4%

For the cement-treated clay, similar to the unreinforced clay, the specimens at higher water content would be failed at a higher axial strain. However, due to the solidification of cement and clay mixture, its UCS increased significantly with the increment of curing days but reduced the axial strain at failure. At 28 days of curing, all of the UCS of cement-treated clay specimens were much higher than that of unreinforced specimens.

### C. The influence of water content on the development of UCS of cement-treated clay

As shown in Figure 7, the UCS of cement-treated clay specimens was increased consistently with the curing days. After 28 curing days, the UCS of the reinforced specimens prepared at the water content in the range of 19.5% - 28.5%, (i.e., around 5% OMC difference) was 225-292 kPa, which was slightly higher than the maximum UCS of the unreinforced specimens. For the specimens at the initial water content of OMC+10% (i.e., 35.5%), their UCS rose sharply from less than 100kPa at the beginning of the curing process to over 600 kPa after 28 curing days. The enormous strength improvement was also observed for specimens with the initial water content of 40.4%, in which the UCS increased from 38.7 kPa to 235.8 kPa when curing in 28 days.

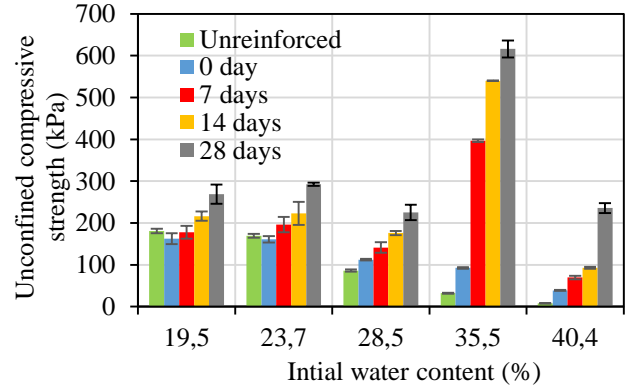


Figure 7. The development of UCS of cement-treated clay specimens at different initial water contents

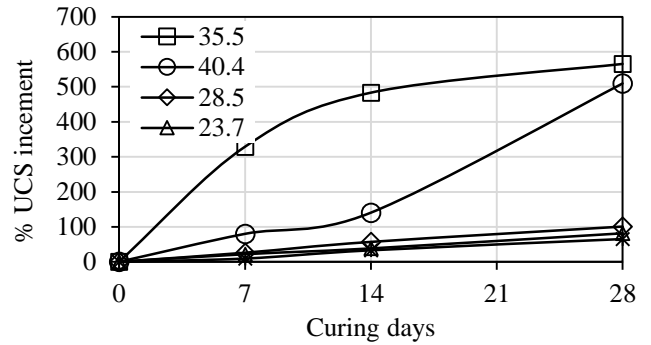


Figure 8. The percentage of UCS increment with curing days of cement-treated clay at different initial water contents

The improvement of UCS of the reinforced specimens with curing days was quantified using the percentage of UCS increment at  $n$  curing days,  $\% \Delta UCS_n$ .

$$\% \Delta UCS_n = \left( \frac{UCS_{re\_n}}{UCS_{re\_0}} - 1 \right) \times 100 \quad (3)$$

in which  $UCS_{re,n}$  and  $UCS_{re,0}$  are the unconfined compressive strength of the reinforced specimens at  $n$  days and 0 days of curing, respectively.

Figure 8 shows that after 28 curing days, the UCS of specimens at OMC $\pm$ 5% would increase 100%, while those of specimens at OMC+10% and OMC+15% were over 500% (Figure 8). The results illustrate that the greatest UCS improvement in the cement-clay mixture required sufficient water content, which was higher than the OMC by 10%. The higher water content of reinforced specimens would induce a decrease in UCS.

## V. CONCLUSION

A series of compaction tests and unconfined compressive tests were performed to investigate UCS and compaction behavior of the cement-treated clay under different water content. The results expressed the importance of controlling the initial water content to achieve the greatest UCS improvement of the cement-treated clay specimens. The following conclusions are drawn from this study.

- The cement binder induced the decrement in OMC and MDD and the optimum degree of saturation of the cement – clay mixture under the standard Proctor compaction test.
- The higher the water content, the more ductile stress-strain behavior was observed. For the unreinforced clay specimens, the maximum UCS occurred at the specimens compacted at the OMC of the clayey soil.
- The UCS of cement-treated clay increased with the days of curing. For the specimens stabilized with 10% cement and compacted at OMC  $\pm$ 5%, their UCS was doubled after 28 curing days. When increasing the initial water content from OMC+10% to OMC+15%, the UCS enhancement of reinforced clay specimens dramatically increased as high as at over 500%. The highest UCS of cement-treated clay was found at the specimens compacted at OMC+10%.

Last, the enormous improvement of the UCS of the compacted clay stabilized by cement approved the potential application of the proposed method for soft clay improvement in reality.

## ACKNOWLEDGMENT

The author sincerely appreciates the constructive comments and feedback by the anonymous reviewers.

## REFERENCES

- [1] J. K. Mitchell, "Soil improvement state of the art report," in *Proc., 10th Int. Conf. on Soil Mechanics and Foundation Engineering*, 4, 509–565., 1981.
- [2] J. N. Meegoda, K. Partymiller, M.K. Richards, W. Kamolpornwijit, W. Librizzi, T. Tate, ... S. Santora, "Remediation of Chromium-Contaminated Soils— Pilot-Scale Investigation," *Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management*, vol. 4, no. 1, pp. 7-15, 2000.
- [3] W. Zhu, C.L. Zhang & A.C.F. Chiu, "Soil–Water Transfer Mechanism for Solidified Dredged Materials," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 133, no. 5, p. 588–598, 2007.
- [4] B. R. Phanikumar & E. Ramanjaneya Raju, "Compaction and strength characteristics of an expansive clay stabilised with lime sludge and cement," *Soils and Foundations*, vol. 60, no. 1, pp. 129–138, 2020.
- [5] W. Zhu, Y. H. Huang, C. L. Zhang & Q. S. Liu., "Effect of curing time on mechanical behavior of crushed solidified dredged material," *Characterization, Monitoring, and Modeling of Geosystems (GSP179)*, vol. ASCE 179, p. 597–604, 2008.
- [6] Y. Huang, W. Zhu, X. Qian, N. Zhang, & X. Zhou, "Change of mechanical behavior between solidified and remolded solidified dredged materials," *Engineering Geology*, vol. 119, no. 3-4, p. 112–119, 2011.
- [7] S. C. Chian, S.T. Nguyen & K.K. Phoon, "Extended Strength Development Model of Cement-Treated Clay," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 142, no. 2, p. 06015014, 2016.
- [8] R. Bahar, M. Benazzoug & S. Kenai, "Performance of compacted cement-stabilised soil," *Cement and Concrete Composites*, vol. 26, no. 7, p. 811–820, 2004.
- [9] I.A. Oyediran & M. Kalejaiye, "Effect of Increasing Cement Content on the strength and compaction parameters of some SW lateritic soils," *EJGE*, vol. 26, no. Bund k, pp. 1501-1514, 2011.
- [10] J. K. Mitchell & K. Soga, "Fundamentals of Soil Behavior, 3rd Edition," John Wiley and Sons, 2005, p. 577.
- [11] D. Ribeiro, R. Néri & R. Cardoso, "Influence of Water Content in the UCS of Soil-Cement Mixtures for Different Cement Dosages," *Procedia Engineering*, vol. 143, pp. 59-66, 2016.
- [12] A. D2166, "Standard Test Method for Unconfined Compressive Strength of Cohesive Soil," in *ASTM International*, West Conshohocken, PA, USA.
- [13] I. Shooashpasha & R.A. Shirvani, "Effect of cement stabilization on geotechnical properties of sandy soils," *Geomechanics and Engineering*, vol. 8, no. 1, p. 17–31, 2015.
- [14] G.A. Miller and S. Azad , "Influence of soil type on stabilization with cement kiln dust," *Constr. Build. Mater.*, vol. 14, no. 2, pp. 89-97, 2000.
- [15] F. Sariosseiri & B. Muhunthan, "Effect of cement treatment on geotechnical properties of some Washington States soils," *Eng. Geol.*, vol. 104, no. 1-2, pp. 119-125, 2009.
- [16] M.D. Nguyen, K.H. Yang, and W.M. Yalew, "Compaction Behavior of Nonwoven Geotextile-Reinforced Clay," *Geosynthetics International*, vol. 27, no. 1, pp. 16-33, 2020.
- [17] F. Tatsuoka & A.G. Correia, "Importance of controlling the degree of saturation in soil compaction," *Procedia Engineering*, vol. 143, p. 556–565., 2016.



# Traffic Flow Estimation Using Deep Learning

Tran Nhat Huy  
*Department of Mechatronics,  
 HCMC University of Technology and Education*  
 Ho Chi Minh City, Vietnam  
 trannhathuy.cdt11.spkt@gmail.com

Bui Ha Duc  
*Department of Mechatronics,  
 HCMC University of Technology and Education*  
 Ho Chi Minh City, Vietnam  
 ducbh@hcmute.edu.vn

**Abstract**—In Vietnam, traffic is always a complex and challenging problem due to a mixture of different types of vehicle as well as the large number of vehicles on road. To improve the traffic management, it is critical to develop a real time traffic flow estimation system which can detect, classify and count vehicles, detect traffic violation at any given time. In this study, a multi-vehicle detection and tracking approach was proposed to achieve these requirements. The proposed model involved two major steps: detection and multiple-object tracking. In the first step, the vehicles were detected and classified into classes (motorbike, car, truck, bus and rudimentary vehicles) using Faster R-CNN model. Next, the movement of the detected objects was tracked with CSRT tracker. Then, all vehicle data is sent to an analyzer to estimate the traffic flow by counting and classifying vehicles by their driving direction. Results showed that the model can robustly work in realtime with an accuracy >86%.

**Keywords**— *Traffic, Estimation, Deep Learning, Computer Vision, Vehicles, Faster-RCNN, CSRT tracker.*

## I. INTRODUCTION

Vietnam is one of the fastest growing economies in Asia. The economic growth leads to many benefits such as higher incomes, better living conditions, better education. However, it also poses new set of problems for the authorities. The economic development is not evenly spread across of the provinces with a large portion of population which is concentrated in big cities, leading to a tremendous increase in the number of vehicles and traffic congestion. This rising is inescapable and challenging in the modern cities, but could be handled with a good strategy of traffic management. Recently, many countries have realized the important role of traffic data and started collecting traffic data to develop intelligent transportation systems. Besides, traffic data not only enable political representatives to make more informed decisions but also inform the public about the traffic situation.

To collect traffic data, approaches – both fixed or mobile – have been proposed and applied in practice. The most popular method is to detect passing vehicles at given locations such as tool station, traffic control centers using fixed detector (e.g. inductive loops, pneumatic road tubes or piezoelectric sensors) placed on or in the road [1]. These traditional on-road sensors can register each bypassing vehicle and send raw data to their data centers in real time. However, these methods are suffered from limited coverage and high implementation and maintenance cost. Alternatively, traffic data can also be anonymously collected through mobile phones and in-vehicles GPS [2,3]. This method can provide accurate information about car location, speed and direction of travel. But it requires every vehicle is equipped with mobile phone or GPS, which may be not feasible for the current situation in Vietnam.

In recent year, along with the development of image processing and deep learning techniques, more attention has been paid to the remote observation using video image from traffic camera. This method is very promising since it not only can provide anonymous data of traffic, but also can record vehicle number, type, speed and traffic violation. In 2017, there were 500 million surveillance cameras worldwide, producing billion gigabytes of data weekly and this number doubles every 2 years. In Ho Chi Minh city, there are more than 600 traffic cameras placed at main streets for observing traffic flow. The information obtained by these cameras is enormous. Many studies, using different video processing techniques, have been conducted to analyze these data [4-6]. However, these studies were mostly conducted in developed countries, which have majority of vehicles are cars. In Vietnam, the situation is more complicated due to the mixture of different types of vehicle as well as the large number of vehicles on road.

This study proposes an approach to investigate the complex traffic situation in Vietnam. The proposed model involved two parts: multiple-vehicle detection with Faster R-CNN and online tracking with CSRT algorithm.

## II. METHOD OF TRAFFIC FLOW ESTIMATION

### A. Traffic dataset

A good dataset has crucial role in the performance of deep learning network. Currently, which is available for the traffic estimation such as KITTY, PDTV, MIT, but majority of vehicles in these datasets are cars, which are not sufficient to reflect the traffic in Vietnam. Therefore, in this study a new traffic dataset was collected from surveillance cameras installed in Ho Chi Minh city. This dataset has more than 4000 images, collected in different light and weather condition. The type and location of vehicles in each image are manually labeled using Labelling tool to create the ground truth for the training and testing sessions.

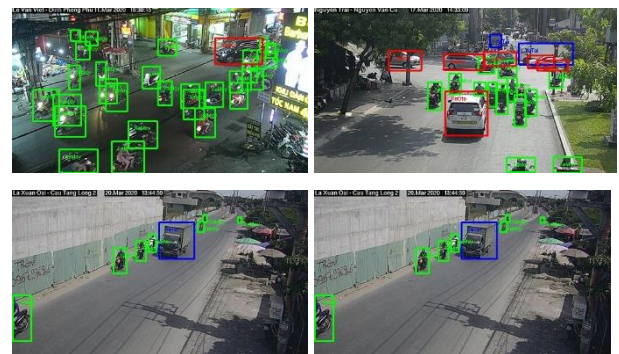


Figure 1. Examples of the new dataset with ground truth

### B. Vehicle detection and classification with Faster R-CNN network

Vehicle detection and classification are the first step in traffic flow estimation. The accurate detection of moving vehicles in a scene is challenging due to the number of occurrences of the objects and occlusion. This problem is even more severe for the traffic situation in Vietnam, where many people use motorcycle on road. Additionally, the detection system need to work in a timely manner so that it can be useful for real time traffic management. Hence, traditional object detection techniques such as SVM, standard convolutional network are not suitable since they are time consuming and require high computational cost. In this study, Faster R-CNN algorithm [7] is used to detect and classify vehicles thanks to its speed and accuracy.

The Faster R-CNN has two networks which shares convolutional features: Region Proposals Network (RPN) receives images at any size and rapidly generate region proposals, and a deep network exam these proposals to check the occurrence of objects. The sharing features enable the detection system run at near real-time speed.

In this paper, an alternating training technique [7] is adopted to train the RPN and the classifier. Both networks were firstly initialized with pre-trained Inception Resnet v2. Next, the RPN was trained so that it can generate 300 bounding boxes with the intersection over union overlap higher than 0.7 compared to ground truth boxes. The anchor boxes of RPN were select at three aspect ratios [1:2, 1:1, 2:1] to match with the ratio of height and width of the vehicles in VietNam. After that, the classifier was trained using the proposal bounding boxes generated in the previous step. The classifier parameters were then used to fine-tune RPN.

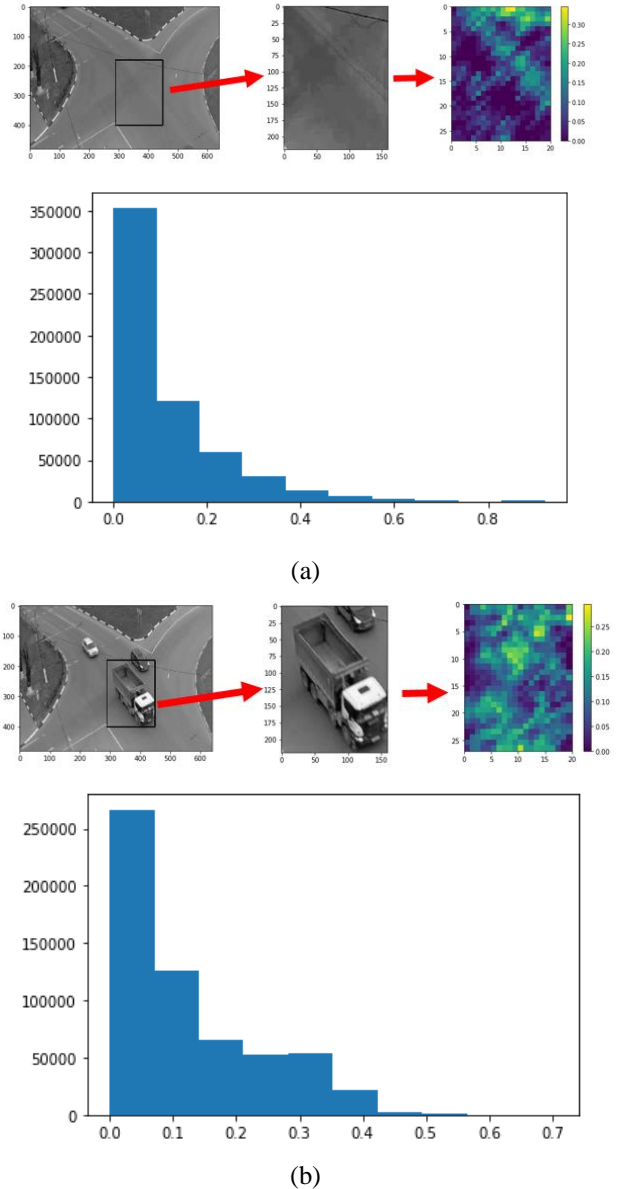


Figure 2. Examples of multiple-vehicle detection with Faster R-CNN

### C. Integrating detection with tracking

To estimate the traffic flow, the system need to be more sophisticated than just detect and classify vehicles. In this

paper, a CSRT tracking algorithm was employed between the detections to track and label vehicles on road. The CSRT calculates histogram of oriented gradients (HoGs) of objects detected in previous frame (Figure 3) and search the area around its last known position in successive frames to predict vehicles locations. The advantages of CSRT tracker can robustly track unpredictable motion, which is common in traffic, and tolerate intermittent frame drops. Additionally, it can adjust the size of target window dynamically [8].



Figures 3. Normalized HoGs of (a) road without vehicles and (b) vehicle

Using tracking techniques helps significantly to reduce the computational cost and suppress false detection. Furthermore, a tracker can also provide movement data of the vehicle to judge driver behavior such as driving speed and lane usage.

### D. Traffic flow estimation

The results of detection and tracking system was used to evaluate the number of traffic parameters such as: number of vehicles on road, its direction, speed.

- Road traffic counting

Vehicles in a frame will be firstly detected with Faster R-CNN. These vehicles were then labelled and tracked using

CSRT algorithm. The bounding box's center of detected vehicles will be continuously monitored to check its location in successive frames.

In highway scenes, a line is used to separate image into 2 parts as shown in Figure 4. If the bounding box's centers pass the line, the number of vehicles entering the road is count up.



Figure 4. Traffic counting at highway scene

In crossroad scene, 4 lines were used to separate image into 5 parts (Figure 5). Each detected vehicle in the frame will be assign a vector  $[x,y]$ , where  $x$  is the entrance part and  $y$  is the exit part. If vehicle entered the road from part 1,  $x=1$ , if vehicle initially belong to part 2,  $x=2$  and so on. Then the vehicle's location was tracked to check where it goes. If the car enters from part 1 and go to part 3, the vector will be  $[1,3]$  (Figure 6). Based on that, we can identify direction of the vehicle at crossroad.

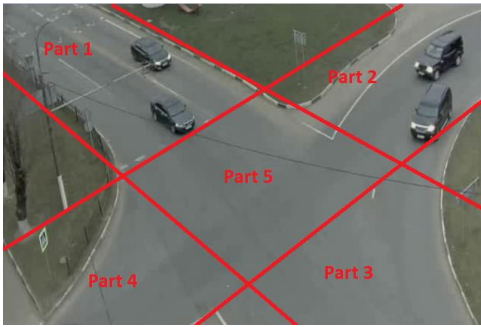


Figure 5. Traffic counting at crossroad scene

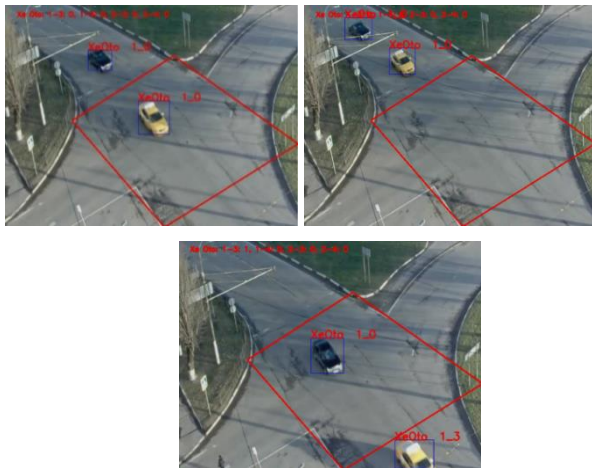


Figure 6. Traffic flow estimation at crossroad

- Estimate vehicle speed

To estimate the speed, the distance that a vehicle has travel during a 10 frame period were measured.

$$v = \frac{d_t - d_{t-10}}{10/fps} \quad (1)$$

where fps is the video rate (frames per second).

Due to the need of marker object for the distance estimation, the vehicle's speed can only measure at few locations.

### III. EXPERIMENTAL RESULTS AND DISCUSSION

#### A. Accuracy of the detection and classification system

The performance of the proposed method was evaluated on traffic scenes at highway and crossroad. Each video clip is 1-minute long, 22 frames per second, captured in both sufficient light and low light conditions. The total number of vehicles of each type were manually counted to validate the automatic counting system.

In sufficient light condition, the results in Figure 7 and Table I showed that the system can detect and track most of the vehicle available in the highway scene and crossroad scene with high accuracy (>90%). The poor performance in detecting rudimentary vehicles could be due to the small size of training data for this class. At crossroad scene, the accuracy drop may be explained by the unpredictable movement of vehicle. When a vehicle changes its direction at cross road, it is quickly covered by other cars, motorbike, makes the tracker becomes less effective.

TABLE I. TRAFFIC FLOW ESIMATION  
IN SUFFICIENT LIGHT CONDITIONS

Type of vehicles	Actual	Counted	Accuracy (%)
<b>Highway scene</b>			
Car	36	35	97.2
Motorbike	212	204	96.2
Pickup truck	10	9	90
Rudimentary vehicles	4	1	25
<b>Crossroad scene</b>			
Car	29	25	86.2
Motorbike	172	152	88.4
Pickup truck	3	3	100
Bus	1	1	100

TABLE II. TRAFFIC FLOW ESIMATION  
IN LOW LIGHT CONDITIONS

Type of vehicles	Actual	Counted	Accuracy (%)
<b>Highway scene</b>			
Car	32	31	96.8
Motorbike	176	170	96.6
Pickup truck	2	2	100
<b>Crossroad scene</b>			
Car	78	74	94.8
Motorbike	132	123	93.1
Bus	6	6	100



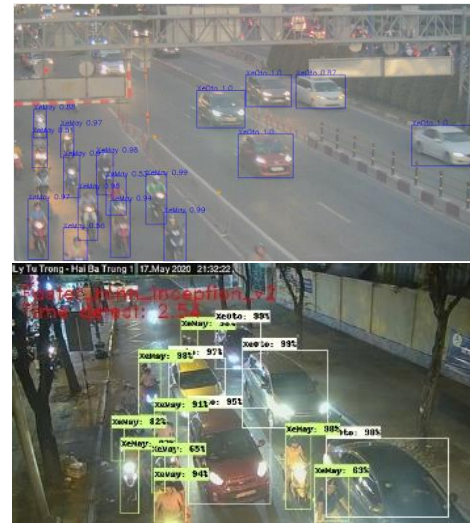


(a)

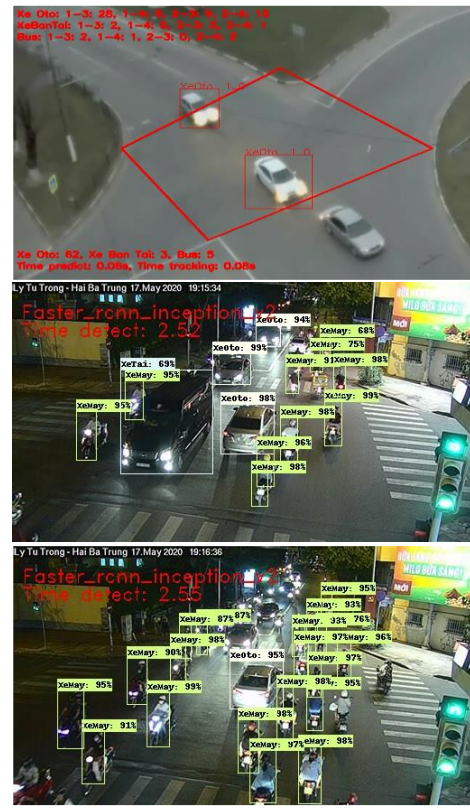


(b)

Figure 7. Vehicles detection in sufficient lighting conditions at (a) highway and (b) crossroad



(a)



(b)

Figure 8. Vehicles detection in low lighting conditions at (a) highway and (b) crossroad

In low light conditions, although the light from vehicles' head lamp cause significant changes in color and pixel values, the system can still keep its performance (Figure 8 and Table 2). It showed that with a good training dataset, the Faster R-CNN and CSRT can be work well even when the light conditions are not sufficient.

#### IV. CONCLUSIONS

In this paper, we proposed an effective model to detect, classify vehicles and estimate traffic flow by counting the number of vehicles on road, track its direction and speed. The result of this paper could be a premise for building a smart system which can management traffic in real time. Although

there were still some missing in detecting, tracking vehicles, the accuracy of the system was acceptable.

In the future, we will continue to improve the dataset and algorithms to increase accuracy, reduce time prediction in tracking.

#### REFERENCES

- [1] Martin, P.T., Feng, Y., Wang, X., Detector Technology Evaluation, Technical Report, Utah Transportation Center, 2003.
- [2] Astarita, V., Bertini, R. L., d'Elia, S., Guido, G., Motorway traffic parameter estimation from mobile phone counts, *European Journal of Operational Research*, Vol.175, pp.1435-1446, 2006.
- [3] Bar-Gera, H., Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel, *Transportation Research Part C*, Vol.15, pp.380-391, 2007.
- [4] Kim, S.H., Shi, J., Alfarrarjeh, A., Xu, D., Tan, Y., Shahabi, C. Real-time traffic video analysis using intel viewmont coprocessor, in: *International Workshop on Databases in Networked Information Systems*, pp. 150–160, 2013.
- [5] Zapletal, D., Herout, A. Vehicle re-identification for automatic video traffic surveillance, in: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1568–1574, 2016.
- [6] Zhuo, L., Jiang, L., Zhu, Z., Li, J., Zhang, J., Long, H., 2017. Vehicle classification for large-scale traffic surveillance videos using convolutional neural networks. *Machine Vision and Applications* 28, 793–802.
- [7] Shaoqing Ren\*, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection. *Microsoft Research*, pp.1-8, 2016.
- [8] Alan Lukezi, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas and Matej Kristan. Discriminative Correlation Filter Tracker with Channel and Spatial Reliability, *International Journal of Computer Vision* volume 126, pp.671–688, 2018.



# A Finite-Time Robust Control for a Manipulator with Output Constraints and Unknown Control Directions

Duc-Thien Tran

Department of Automatic Control,  
Faculty of Electrical and Electronic  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thientd@hcmute.edu.vn

Manh-Son Tran

Department of Automatic Control,  
Faculty of Electrical and Electronic  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
sontm@hcmute.edu.vn

Nguyen Van Hiep

Department of Electronics-Biomedical  
Engineering  
Faculty of Electrical and Electronic  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
hiepspkt@hcmute.edu.vn

**Abstract**—This paper presents a finite time robust control for a robotic manipulator regardless of the time-varying output constraints and unknown control direction. First, a transformation technique is investigated to handle the output constraints by converting a constrained system into a free constrained system. When the stability of the free constrained system is guaranteed, the constrained system will overcome the violation of system output. Next, the backstepping control is developed on the unconstrained system with fractional order term to improve the accuracy and transient time of the output responses. Furthermore, a Nussbaum gain function is integrated into the control algorithm to address the problem of unknown control direction. A numerical simulation is provided to confirm the superiorities of the proposed control.

**Keywords**—manipulator, transformation technique, Nussbaum gain function, backstepping control, output constraints, unknown control direction, finite-time control.

## I. INTRODUCTION

Nowadays, robots have been widely investigated in many applications in industry, surgical operation, military, etc. The conventional issues [1] which are high nonlinearity, coupling torques, unknown friction, and modeling error are challenges in the control aspect. In order to overcome them, many advanced algorithms such as computed torque control [2], adaptive sliding mode control [3-5], adaptive backstepping control [6, 7] have been provided. In recent years, the robot is developed to the next generation, which can co-work with humans in polishing, deburring, assembly, etc. Additionally, new problems such as input constraints, output constraints, etc. also arise. The output constraints are limitations of the output responses, which are generated by the trajectories and environmental information. The role of the output constraints is guaranteed that the accuracy requirements strictly obtain in regions that are defined by the output constraints.

In order to avoid the transgression of the output constraints, some advanced techniques have been developed comprising of model predictive control (MPC) [8], barrier Lyapunov function (BLF) [9], and transforming technique [10] from the universities and institutes. The MPC handles the constraints in both the linear and nonlinear systems based on a finite horizon optimization framework [11]. Since the efficiency of the MPC depends on the accuracy of plant

models and processing speed to solve the control problem, the MPC is usually investigated in the process industries where plants operated in low bandwidth. Subsequently, another method to cope with the output constraints is the barrier Lyapunov function. In 2009, this technique was first introduced to cope with the constant output constraint issue [9]. In [12], a BLF was proposed for a nonlinear system with the presence of the output constraints. Because of the complication of the BLFs, it is difficult to conduct and analyze an advanced controller, which ensures the finite-time convergence. In the next method, a transforming method named prescribed performance control was proposed in [13], which can deal with time-varying output constraints by converting the constrained system into the unconstrained system. Conversely, with the BLF method, the transformed system owns potential properties for developing a controller to cope with finite-time convergence problems. Based on the best of the author's knowledge, there has only been few advanced controllers that are developed by the transformed dynamic system to handle this problem.

In this paper, we propose a finite time robust control for a robotic manipulator regardless of the output constraints to guarantee that the output responses track the desired trajectories. The proposed control is developed from the free constrained model which is converted from the constrained model by the transformation technique. Then, a backstepping technique with fractional order terms is employed on the free constrained one to guarantee the stability of the whole system. From these results, we can affirm that the output responses of the robotic manipulator are not broken the output constraints. Next step, the control is processed with a Nussbaum function to deal with the unknown control directions [14] which happened when the operator has the wrong connection in the terminal or actuators are aged. A Lyapunov function which is used to theoretically prove the stability of the whole system is presented in this paper. Finally, some simulations are conducted to verify the effectiveness of the proposed control on 2 DOF manipulator.

The structure of this paper is organized as follows: Problem descriptions such as manipulator dynamics, nonlinear transformation, Nussbaum function, and Preliminaries are presented in Section II; Section III describes the control design procedure and proof of stability Section IV

exhibits the simulation cases to demonstrate the effectiveness of the proposed method; Finally, some discussions and conclusions are expressed in Section V.

## II. PROBLEM DESCRIPTION

### A. Manipulator dynamics

The n-DOF manipulator dynamics are described by [2]

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{G}(\mathbf{q}) + \mathbf{J}^T(\mathbf{q})\mathbf{f} + \boldsymbol{\tau}_{\text{fric}} = \boldsymbol{\tau} \quad (1)$$

where  $\mathbf{q}, \dot{\mathbf{q}}$ , and  $\ddot{\mathbf{q}} \in R^{n \times 1}$  respectively derive position, angular velocity, and angular acceleration vectors of each joint;  $\mathbf{M}(\mathbf{q}) \in R^{n \times n}$  presents the symmetric and positive definite matrix of inertia;  $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) \in R^{n \times n}$  states the Coriolis and Centrifugal term matrix;  $\mathbf{G}(\mathbf{q}) \in R^{n \times 1}$  describes the gravity term;  $\boldsymbol{\tau}$  presents torque acting on joints;  $\mathbf{J}(\mathbf{q})$  is a nonsingular Jacobian matrix;  $\boldsymbol{\tau}_{\text{fric}}$  derives the unknown frictions; and  $\mathbf{f}$  denotes external disturbances.

**Assumption 1:** The external disturbances and unknown friction functions are bounded functions.

The manipulator properties [1] are presented as follows:

**Property 1:**  $\dot{\mathbf{M}}(\mathbf{q}) - 2\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$  is a skew-symmetric matrix, that is given  $\mathbf{x}^T [\dot{\mathbf{M}}(\mathbf{q}) - 2\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})] \mathbf{x} = 0$ .

Let  $\mathbf{x}_1 = \mathbf{q} \in R^n$ , and  $\mathbf{x}_2 = \dot{\mathbf{q}} \in R^n$ , the robotic dynamics (1) can be represented in the state space form as

$$\begin{aligned} \dot{\mathbf{x}}_1(t) &= \mathbf{x}_2(t) \\ \dot{\mathbf{x}}_2(t) &= \mathbf{M}^{-1}(\mathbf{x}_1(t))(\mathbf{u}(t) - \mathbf{C}(\mathbf{x}_1(t), \mathbf{x}_2(t))\mathbf{x}_2(t) - \mathbf{G}(\mathbf{x}_1(t)) - \boldsymbol{\Delta}(t)) \end{aligned} \quad (2)$$

where  $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T, (i=1, 2)$ ;  $\boldsymbol{\Delta} = \mathbf{J}^T(\mathbf{x}_1)\mathbf{f} + \boldsymbol{\tau}_{\text{fric}}$  presents a lumped disturbance, which consists of external disturbance, and unknown friction.  $\mathbf{u}$  presents the input torque  $\boldsymbol{\tau}$ . The control objective is to track a desired trajectory  $\mathbf{x}_d = [x_{d1}, x_{d2}, \dots, x_{dn}]^T$  while guarantees the satisfaction of the system output conditions  $\underline{x}_{li}(t) < x_{li}(t) < \overline{x}_{li}(t)$ , where  $\underline{\mathbf{x}}_l(t) = [\underline{x}_{l1}(t), \dots, \underline{x}_{ln}(t)]^T$  and  $\overline{\mathbf{x}}_l(t) = [\overline{x}_{l1}(t), \dots, \overline{x}_{ln}(t)]^T$  are lower boundary and upper boundary function vector of the system output.

### B. Nussbaum function

A function  $N(\zeta)$  can be defined as a Nussbaum-like function if it has the following properties [15]

$$\limsup_{s \rightarrow \infty} \int_{s_0}^s N(\zeta) d\zeta = +\infty \quad (3)$$

$$\liminf_{s \rightarrow \infty} \int_{s_0}^s N(\zeta) d\zeta = -\infty \quad (4)$$

In this paper, the even Nussbaum function  $N(\zeta) = e^{\zeta^2} \cos\left(\frac{\pi}{2}\zeta\right)$  with the property of  $N(0) = 0$  is taken into account.

**Lemma 1 [16]:** Let  $V(t)$  and  $\zeta(t)$  be smooth functions defined on  $[0, t_f)$  with  $V(t) > 0, \forall t \in [0, t_f)$ . For any  $t \in [0, t_f)$ , if the following inequality holds:

$$V(t) < c_0 + e^{-c_1 t} \int_0^t (g(\tau)N(\zeta) + 1)\dot{\zeta} e^{c_1 \tau} d\tau \quad (5)$$

where constants  $c_i (i=0, 1) > 0$  are suitable positive constants,  $g(\tau)$  is a time-varying parameter which takes values in the unknown closed intervals  $I := [l^-, l^+]$  with  $0 \notin I$ , then  $V(t)$ ,  $\zeta(t)$  and  $\int_0^t g(\tau)N(\zeta)\dot{\zeta} d\tau$  must be bounded on  $[0, t_f)$ .

## III. PROPOSED METHOD

### A. Control description

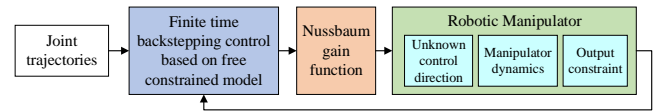


Fig. 1 Diagram of the proposed control

Figure 1 plot the diagram of the proposed control in the robotic manipulator. The robotic manipulator exists two problems which are the output constraints and unknown control directions. In order to overcome these issues, the proposed control is constructed from the finite time backstepping control based on a free constrained model and the Nussbaum gain function. Figure 2 presents the control design procedure of the proposed method. In the first step, a transformation technique is used to convert the constrained system into the unconstrained one. In the next step, the backstepping technique and fractional order term are used to design a finite time backstepping control with the known control direction. Finally, a Nussbaum gain function is applied to develop the proposed control from the controller in step 2.

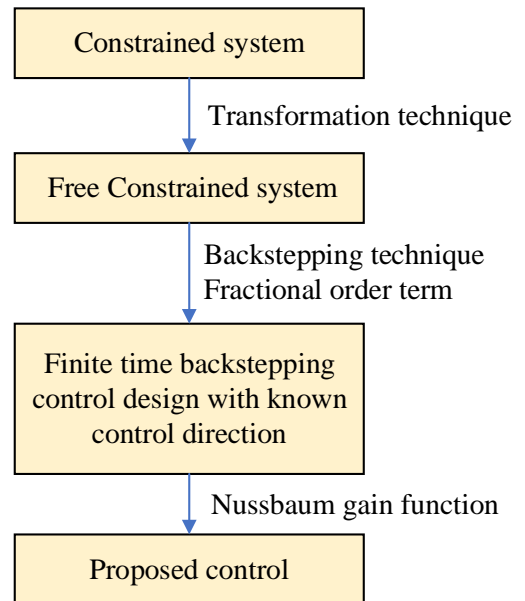


Fig. 2 Control design procedure

### B. Control design

**Step 1:** Convert the constrained system into the unconstrained system.

The error vectors of the system (2) are expressed as follows:

$$\mathbf{e}_1 = \mathbf{x}_1 - \mathbf{x}_d \quad (6)$$

The output errors constraints are bounded the output errors as follows

$$\underline{e}_{li}(t) < e_{li}(t) < \bar{e}_{li}(t) \quad (7)$$

where  $\underline{e}_{li}(t)$  and  $\bar{e}_{li}(t)$  are respectively the lower bound and the upper bound of the error constraints which are defined as follows:

$$\bar{e}_{li}(t) = \bar{x}_{li}(t) - x_{di}(t) > 0; \underline{e}_{li}(t) = x_{li}(t) - x_{di}(t) < 0 \quad (8)$$

where  $\bar{x}_{li}(t)$  and  $\underline{x}_{li}(t)$  are the upper and lower time-varying bounds of the output constraints, respectively.

In order to integrate the errors,  $e_{li}(t)$  with their constraints, a nonlinear transformation scheme is provided to convert the constrained system into an unconstrained one. The transformation equation [17] is written as follows

$$e_{li}(t) = \bar{e}_{li}(t) H_i(s_i(t), \eta_i(t)) \quad (9)$$

where  $\eta_i(t) = \bar{e}_{li}(t) / \underline{e}_{li}(t)$ ,  $s_i$  is the new error variable,  $H_i(\cdot)$  is an increasing and invertible function with respect to  $s_i(t)$ , which satisfies the following conditions

$$\lim_{s_i(t) \rightarrow -\infty} (H_i(s_i(t), \eta_i(t))) = \eta_i(t), \lim_{s_i(t) \rightarrow +\infty} (H_i(s_i(t), \eta_i(t))) = 1 \quad (10)$$

Now, the new error variable,  $s_i(t)$ , can be computed as

$$s_i(t) = H_i^{-1}\left(\frac{e_{li}(t)}{\bar{e}_{li}(t)}, \eta_i(t)\right) \quad (11)$$

**Remark 1:** When the new variable  $s_i(t)$  is bounded, the following inequality holds:

$$\eta_i(t) < H_i(s_i(t), \eta_i(t)) < 1 \quad (12)$$

By replacing (8) and (9) into (12), we can present the inequality as follows:

$$\underline{e}_{li}(t) < \bar{e}_{li}(t) H_i(s_i(t), \eta_i(t)) = e_{li}(t) < \bar{e}_{li}(t) \quad (13)$$

which implied that the output responses are bounded by the output constraints, as shown in (7).

The differentiating  $s_i(t)$  with respect to time is calculated as follows:

$$\dot{s}_i = \frac{\partial H_i^{-1}}{\partial \left(\frac{e_{li}(t)}{\bar{e}_{li}(t)}\right)} \frac{1}{\bar{e}_{li}(t)} \left( \dot{e}_{li}(t) - \frac{e_{li}(t) \dot{\bar{e}}_{li}(t)}{\bar{e}_{li}(t)} \right) + \frac{\partial H_i^{-1}}{\partial \eta_i(t)} \dot{\eta}_i(t) \quad (14)$$

When we replace (14) into the first equation of (2), the manipulator dynamics with output constraint in (2) is transformed into the unconstrained system as follows:

$$\begin{aligned} \dot{s}_1(t) &= \Phi \mathbf{x}_2(t) + \Psi \\ \dot{\mathbf{x}}_2(t) &= \mathbf{M}^{-1}(\mathbf{x}_1(t)) (\mathbf{u}(t) - \mathbf{C}(\mathbf{x}_1(t), \mathbf{x}_2(t)) \mathbf{x}_2(t) \\ &\quad - \mathbf{G}(\mathbf{x}_1(t)) - \Delta(t)) \end{aligned} \quad (15)$$

where

$$\Phi = \begin{bmatrix} \frac{\partial T_1^{-1}}{\partial \left(\frac{e_{11}(t)}{e_{11}(t)}\right)} \frac{1}{e_{11}(t)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\partial T_n^{-1}}{\partial \left(\frac{e_{1n}(t)}{e_{1n}(t)}\right)} \frac{1}{e_{1n}(t)} \end{bmatrix} \in R^{n \times n} \quad (16)$$

$$\Psi = -\mathbf{H} \dot{\mathbf{x}}_d - \begin{bmatrix} \frac{\partial T_1^{-1}}{\partial \left(\frac{e_{11}(t)}{e_{11}(t)}\right)} \frac{1}{e_{11}(t)} \frac{e_{11}(t) \dot{e}_{11}(t)}{e_{11}(t)} \\ \vdots \\ \frac{\partial T_n^{-1}}{\partial \left(\frac{e_{1n}(t)}{e_{1n}(t)}\right)} \frac{1}{e_{1n}(t)} \frac{e_{1n}(t) \dot{e}_{1n}(t)}{e_{1n}(t)} \end{bmatrix} + \begin{bmatrix} \frac{\partial T_1^{-1}}{\partial (\eta_1(t))} \dot{\eta}_1(t) \\ \vdots \\ \frac{\partial T_2^{-1}}{\partial (\eta_2(t))} \dot{\eta}_2(t) \end{bmatrix} \quad (17)$$

**Step 2:** Design a finite time backstepping control with known control direction.

The tracking errors in the unconstrained system (15) are defined

$$\begin{aligned} \mathbf{e}_s &= s_1 \\ \mathbf{e}_2 &= \mathbf{x}_2 - \mathbf{a}_1 \in R^{n \times 1} \end{aligned} \quad (18)$$

where  $\mathbf{a}_1$  is the virtual control vector.

The virtual control is chosen as follows:

$$\mathbf{a}_1 = \Phi^{-1} \left( -\mathbf{K}_{10} \mathbf{e}_s - \mathbf{K}_{11} |\mathbf{e}_s|^{\beta_2} \text{sign}(\mathbf{e}_s) - \Psi \right) \quad (19)$$

where  $\mathbf{K}_{1i} \in R^{n \times n}$  ( $i=0,1$ ) present positive diagonal matrices;  $0 < \beta_2 < 1$  is a positive constant.

The control law is developed as follows:

$$\begin{aligned} \mathbf{u}(t) &= -\Phi \mathbf{e}_s - \mathbf{K}_{20} \mathbf{e}_2 - \mathbf{K}_{21} |\mathbf{e}_2|^{\beta_2} \text{sign}(\mathbf{e}_2) + \mathbf{C}_0(\mathbf{x}_1, \mathbf{x}_2) \mathbf{a}_1 \\ &\quad + \mathbf{G}_0(\mathbf{x}_1) + \mathbf{M} \dot{\mathbf{a}}_1 \end{aligned} \quad (20)$$

where  $\mathbf{K}_{2i} \in R^{n \times n}$  ( $i=0,1$ ) presents a positive diagonal matrix.

**Theorem 1:** When the control laws in (19) and (20) are applied for the manipulator dynamics (15), the finite-time stability of the controlled system is guaranteed, and the output constraints are satisfied. The residual set of the manipulator dynamics is given by

$$\lim_{t \rightarrow T_r} |V(e)| \leq \min \left\{ \frac{\delta}{(1-\phi_0)\lambda_1}, \left( \frac{\delta}{(1-\phi_0)\lambda_2} \right)^{\frac{2}{1+\beta_2}} \right\} \quad (21)$$

where  $0 < \phi_0 < 1$ ;  $\mathbf{e} = [\mathbf{e}_s^T \quad \mathbf{e}_2^T]^T$ . The finite time is

$$T_r \leq \max \left\{ t_0 + \frac{2}{\phi_0(1-\beta_2)} \ln \frac{\phi_0 \kappa_1 V^{\frac{1+\beta_2}{2}}(e(t_0)) + \kappa_2}{\kappa_2}, t_0 + \frac{2}{\kappa_1(1-\beta_2)} \ln \frac{\kappa_1 V^{1-\beta_2}(e(t_0)) + \phi_0 \kappa_2}{\phi_0 \kappa_2} \right\} \quad (22)$$

**Remark 2:** In order to prove the stability of Theorem 1, we derive a Lyapunov function as follows:

$$V_2 = \frac{1}{2} \mathbf{e}_s^T \mathbf{e}_s + \frac{1}{2} \mathbf{e}_2^T \mathbf{M} \mathbf{e}_2 \quad (23)$$

with Lemma 2 and Lemma 3. The proof of results will be discussed in future work.

**Step 3:** Using the Nussbaum function to design the proposed control with the unknown control directions.

The unconstrained dynamics are represented to exhibit the unknown control directions as follows

$$\begin{aligned} \dot{\mathbf{s}}_1(t) &= \Phi \mathbf{x}_2(t) + \Psi \\ \dot{\mathbf{x}}_2(t) &= \mathbf{M}^{-1}(\mathbf{x}_1(t)) (\tilde{\mathbf{K}} \mathbf{u}(t) - \mathbf{C}(\mathbf{x}_1(t), \mathbf{x}_2(t)) \mathbf{x}_2(t) \\ &\quad - \mathbf{G}(\mathbf{x}_1(t)) - \Delta(t)) \end{aligned} \quad (24)$$

where  $\tilde{\mathbf{K}}$  depict the unknown control directions.

The proposed control laws are defined as follows:

$$\begin{aligned} \mathbf{u}(t) &= N(\zeta) \left( -\Phi \mathbf{e}_s - \mathbf{K}_{20} \mathbf{e}_2 - \mathbf{K}_{21} |\mathbf{e}_2|^{\beta_2} \text{sign}(\mathbf{e}_2) \right. \\ &\quad \left. + \mathbf{C}_0(\mathbf{x}_1, \mathbf{x}_2) \mathbf{a}_1 + \mathbf{G}_0(\mathbf{x}_1) + \mathbf{M} \dot{\mathbf{a}}_1 \right) \end{aligned} \quad (25)$$

where  $N(\zeta) = e^{\zeta^2} \cos\left(\frac{\pi}{2} \zeta\right)$ .

$$\begin{aligned} \dot{\zeta} &= -\Phi \mathbf{e}_s - \mathbf{K}_{20} \mathbf{e}_2 - \mathbf{K}_{21} |\mathbf{e}_2|^{\beta_2} \text{sign}(\mathbf{e}_2) \\ &\quad + \mathbf{C}_0(\mathbf{x}_1, \mathbf{x}_2) \mathbf{a}_1 + \mathbf{G}_0(\mathbf{x}_1) + \mathbf{M} \dot{\mathbf{a}}_1 \end{aligned} \quad (26)$$

**Remark 3:** In order to prove the stability of the whole system with the proposed control, we analysis the Lyapunov function with Lemma 1. This analysis will conduct in future work.

#### IV. SIMULATION

##### A. Simulation descriptions

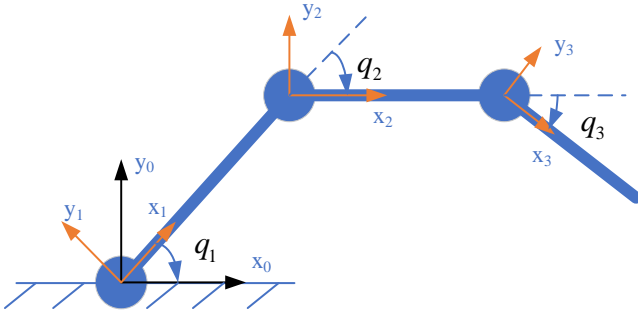


Fig. 3 Structure of 3-DOF manipulator

Some simulations are conducted on MATLAB Simulink with a three DOF manipulator to illustrate the superiorities of the proposed control. The MATLAB Simulink is configured with the sampling time of 0.01 (second); Solver type: ODE3. Additionally, the simulation time is 40 seconds.

The 3-DOF manipulator presented in Fig. 3 is a planar robot with 3 rotary actuators. The parameters of the manipulator are presented in Table 1. Additionally, all mass exists as a point mass at the distal end of each link, the center of mass in each link is presented by  ${}^i P_C = l_i X_i, (i = 1, 2, 3)$ .

TABLE 1. PARAMETERS OF THE 3-DOF MANIPULATOR

Symbol	Description	Symbol	Description
$l_1 = 0.35m$	Length of 1 <sup>st</sup> link.	$m_1 = 0.23kg$	Mass of 1 <sup>st</sup> link
$l_2 = 0.3m$	Length of 2 <sup>nd</sup> link.	$m_2 = 0.2kg$	Mass of 2 <sup>nd</sup> link
$l_3 = 0.15m$	Length of 3 <sup>rd</sup> link.	$m_3 = 0.1kg$	Mass of 3 <sup>rd</sup> link
$g = 9.81ms^{-2}$	Gravity constant		

By using the Newton iteration method in [2], the dynamics of the manipulator are presented as follows:

$$\mathbf{M}(\mathbf{q}) \ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) \dot{\mathbf{q}} + \mathbf{G}(\mathbf{q}) + \mathbf{J}^T(\mathbf{q}) \mathbf{f} + \boldsymbol{\tau}_{fric} = \boldsymbol{\tau} \quad (27)$$

The friction model vector includes the viscous and coulomb frictions, which is presented as follows:

$$\boldsymbol{\tau}_{fric} = b \dot{\mathbf{q}} + c \text{sign}(\dot{\mathbf{q}}) \in R^3 \quad (28)$$

where  $b = 0.5 \text{diag}([1, 1, 1])(Nms / \text{rad})$ ,  $c = \text{diag}([1, 1, 1])(Nm)$

During the simulation period, an external disturbance along the x-axis of the origin Coordinator is applied after 20<sup>th</sup> second,  $\mathbf{f} = -50x_0(N)$

The trajectory signals,  $x_d$  and  $z_d$ , in the Cartesian coordinator are sine waves,  $x_d = 0.4 + 0.15 \cos(0.2\pi t)(m)$ ,  $y_d = 0(m)$ , and  $z_d = 0.15 \sin(0.2\pi t)(m)$ . Additionally, the rotary angle around the z-axis is zero.

##### B. Simulation results

The merits of the proposed controller are illustrated through comparisons with backstepping control (BC) and backstepping control (BC) with the new transformed model (20) and (21).

The parameters of these controllers are described in Table 2.

TABLE 2. THE PARAMETERS OF THREE CONTROLLERS

Controllers	Parameters
BC	$K_{10} = 10 \text{diag}([4, 4, 3])$ , $K_{20} = 0.5 \text{diag}(3, 3, 1)$
BC with output constraints	$K_{10} = 10 \text{diag}([4, 4, 3])$ , $K_{20} = 0.5 \text{diag}(3, 3, 1)$ $e_{li} = \begin{cases} 2 & t \leq 5 \\ 2e^{-0.4(t-5)} + 0.04 & \text{otherwise} \end{cases}, \quad \underline{e}_{li} = -\overline{e}_{li}$
Proposed controller	$K_{10} = 2 \text{diag}([4, 4, 3])$ , $K_{11} = 8 \text{diag}([4, 4, 3])$ $K_{20} = 0.1 \text{diag}(3, 3, 1)$ , $K_{21} = 0.4 \text{diag}(3, 3, 1)$ $e_{li} = \begin{cases} 2 & t \leq 5 \\ 2e^{-0.4(t-5)} + 0.04 & \text{otherwise} \end{cases}, \quad \underline{e}_{li} = -\overline{e}_{li}$

Fig. 4 shows the output responses of all joints in the manipulator with the output constraints. Three controllers are designed how to guarantee stability under the existence of the frictions and external disturbance. It is difficult to realize the advantages of the proposed controllers in Fig. 4. Then Fig. 5 presented the output response errors and output constraint error. The results in this figure exhibit that the errors of three

controllers are not eliminated because of the unknown friction and external disturbance. The responses of the BC break the output constraints since the output constraint issues are not taken into account during control design. On the contrary, the BC with output constraints and the proposed control is designed with the output constraints, so the output responses avoid the transgression of the output constraints. Fig. 6 shows the control signals of the proposed control.

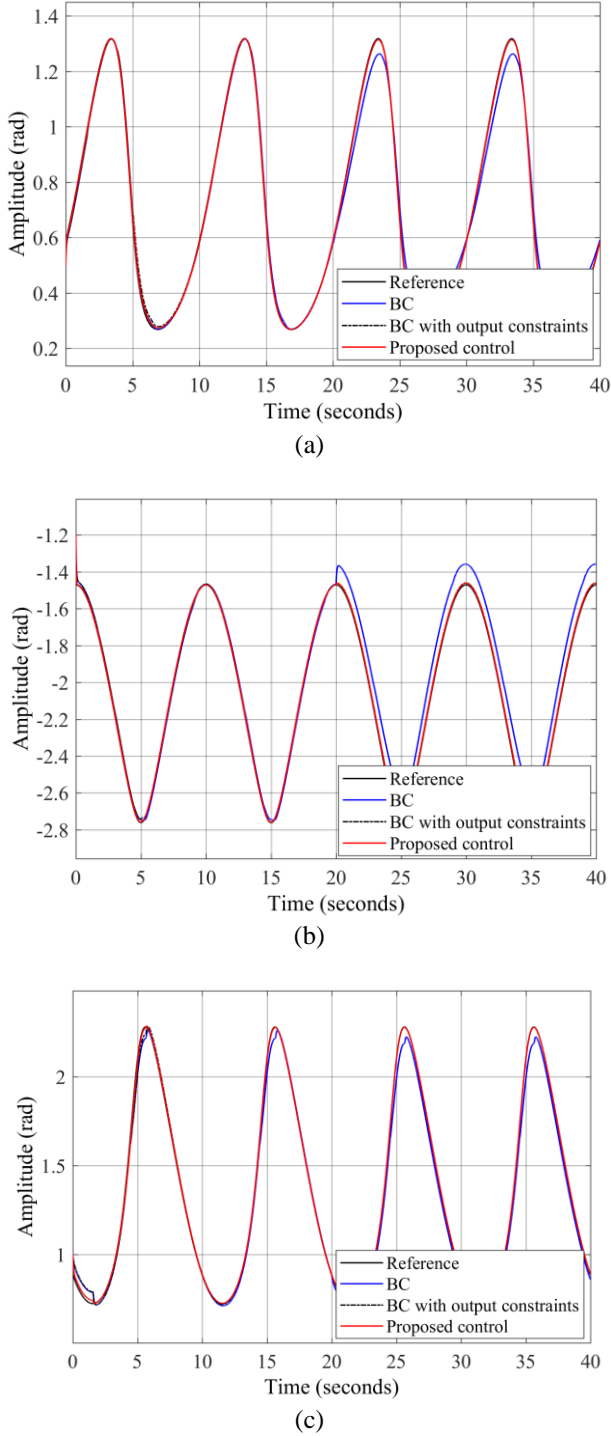


Fig. 4 Output responses of the 3-DOF manipulator with three controllers in a) joint 1, b) joint 2, and c) joint 3

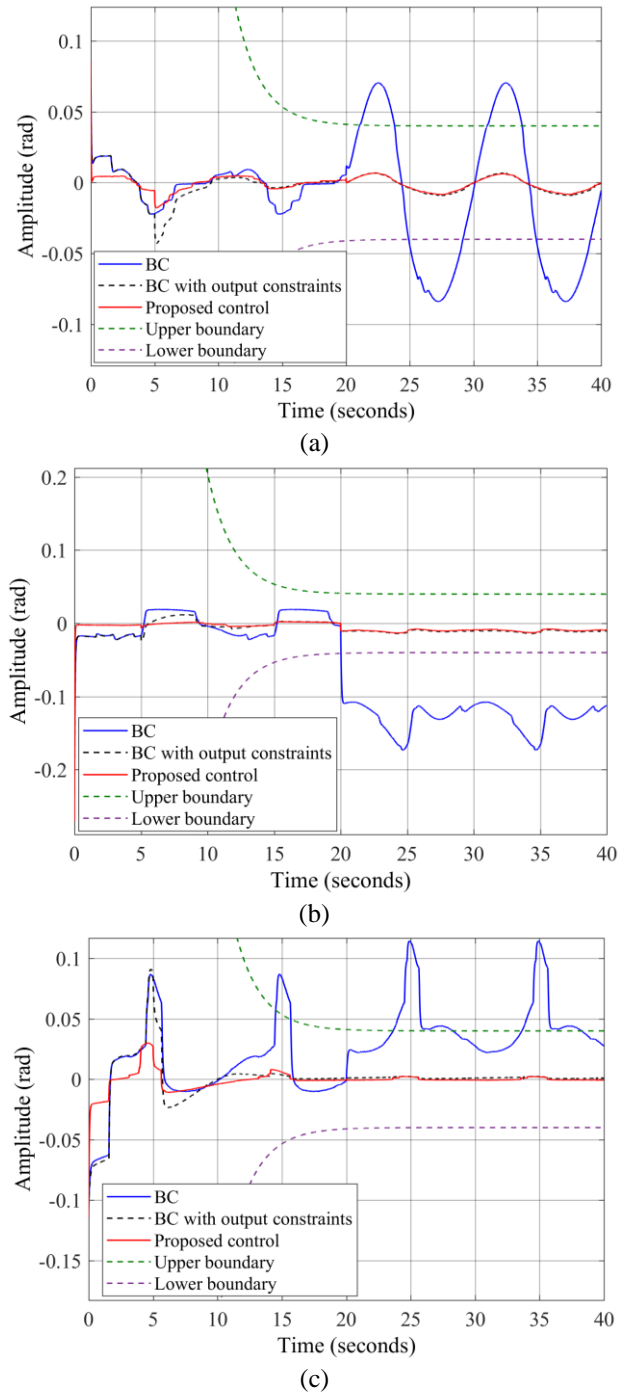


Fig. 5 Error responses of the 3-DOF manipulator with the proposed control

Next simulation, the unknown control direction is considered with the same trajectory inputs. The gain input of the manipulator is set up  $\tilde{\kappa} = \text{diag}([1, 1, -0.2])$ . Fig. 7 shows the response of the Nussbaum gain function with the gain input of the manipulator. The Nussbaum gain is adjusted to compensate with the control laws to solve the unknown control direction. Finally, the control signals in Fig. 8 are the torque responses of the proposed control in this case.



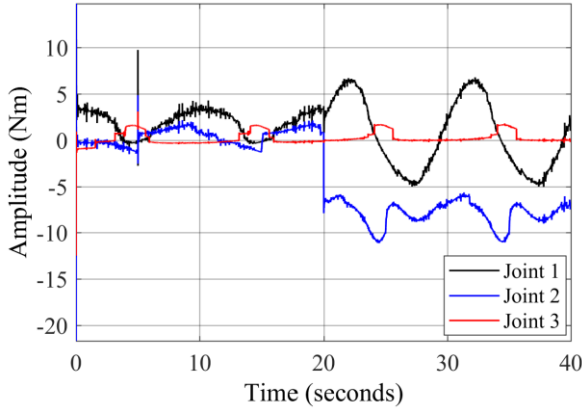


Fig. 6 Torque responses of the proposed control

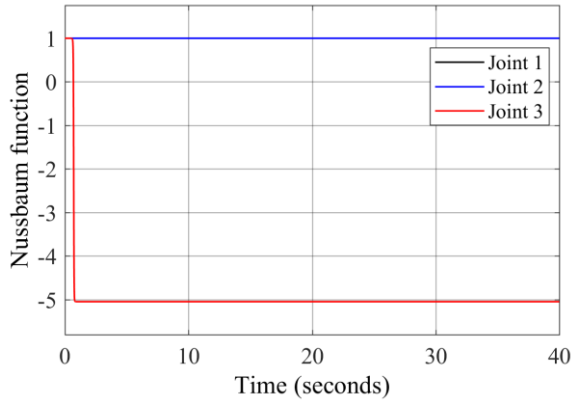


Fig. 7 Nussbaum function responses with wrong control direction in joint 3

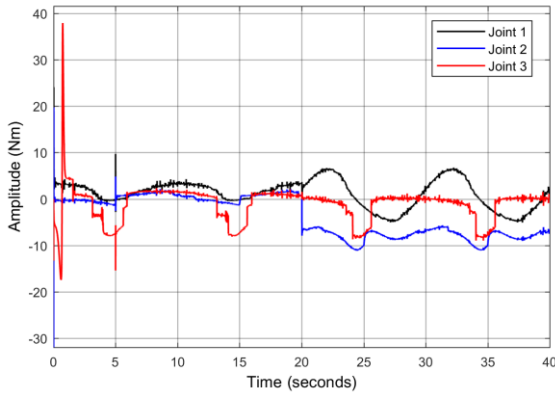


Fig. 8 Torque responses of the proposed control

## V. CONCLUSION

This paper proposed a finite time robust control for an n-DOF manipulator regardless of unknown control direction and output constraints. The control design is divided into three steps. In the first step, a transformation technique is applied to the constrained system to convert it into the unconstrained one. Based on this result, a finite time backstepping control is developed to guarantee the stability of the whole system. When all state errors of the unconstrained one are bounded, the stability and output constraint satisfaction in the constrained one are achieved. Next, a Nussbaum gain function is integrated in the finite time backstepping control to design the proposed control. These results help the proposed control can overcome the unknown control direction. Lyapunov functions for analyzing the stability of the system is introduced in this paper. Finally, some simulations are

conducted to validate the effectiveness of the proposed control.

Future work, the stability of the whole system with the proposed control will be investigated and some experiments will be conducted to show the advantages of the proposed method.

## APPENDIX

The position of the end-effector in the 3-DOF manipulator is computed as follows:

$$P_x = l_1 c_1 + l_2 c_{12} + l_3 c_{123}; P_z = l_1 s_1 + l_2 s_{12} + l_3 s_{123}$$

The Jacobian matrix of the robot is expressed by

$$J = \begin{bmatrix} -l_1 s_1 - l_2 s_{12} - l_3 s_{123} & -l_2 s_{12} - l_3 s_{123} & -l_3 s_{123} \\ 0 & 0 & 0 \\ l_1 c_1 + l_2 c_{12} + l_3 c_{123} & l_2 c_{12} + l_3 c_{123} & l_3 c_{123} \end{bmatrix}$$

The elements of the inertia matrix, Coriolis matrix, and gravity matrix are calculated as follows:

$$M(q) = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}, C(q, \dot{q}) = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}, G(q) = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix}$$

$$\begin{aligned} M_{11} &= l_2 (m_3 s_3 (s_3 (l_2 + l_1 c_2) + l_1 c_3 s_2) \\ &\quad + m_3 c_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3)) \\ &\quad + l_1 (s_2 (l_1 m_2 s_2 + m_3 c_3 (s_3 (l_2 + l_1 c_2) + l_1 c_3 s_2) \\ &\quad - m_3 s_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3)) \\ &\quad + c_2 (m_2 (l_2 + l_1 c_2) + m_3 s_3 (s_3 (l_2 + l_1 c_2) + l_1 c_3 s_2) \\ &\quad + m_3 c_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3))) + l_1^2 m_1 \\ &\quad + l_3 m_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3) + l_2 m_2 (l_2 + l_1 c_2) \\ M_{12} &= l_2 (m_3 s_3 (s_3 (l_2 + l_1 c_2) + l_1 c_3 s_2) \\ &\quad + m_3 c_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3)) \\ &\quad + l_3 m_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3) + l_2 m_2 (l_2 + l_1 c_2) \\ M_{13} &= l_3 m_3 (l_3 + c_3 (l_2 + l_1 c_2) - l_1 s_2 s_3); M_{21} = M_{12} \\ M_{22} &= l_2^2 m_2 + l_2 (l_2 m_3 s_3^2 + m_3 c_3 (l_3 + l_2 c_3)) + l_3 m_3 (l_3 + l_2 c_3) \\ M_{23} &= l_3 m_3 (l_3 + l_2 c_3); M_{31} = M_{13}; M_{32} = M_{23}; M_{33} = l_3^2 m_3 \\ C_{11} &= -\dot{x}_2 l_1 l_2 m_3 s_{23} - \dot{x}_3 l_1 l_3 m_3 s_{23} - \dot{x}_2 l_1 l_2 m_3 s_2 - \dot{x}_3 l_2 l_3 m_3 s_3 \\ C_{12} &= -\dot{x}_1 l_1 l_3 m_3 s_{23} - \dot{x}_2 l_1 l_3 m_3 s_{23} - \dot{x}_3 l_1 l_3 m_3 s_{23} - \dot{x}_1 l_1 l_2 (m_2 + m_3) s_2 \\ &\quad - \dot{x}_2 l_1 l_2 (m_2 + m_3) s_2 - \dot{x}_3 l_2 l_3 m_3 s_3 \\ C_{13} &= -l_3 m_3 (l_1 s_{23} + l_2 s_3) (\dot{x}_1 + \dot{x}_2 + \dot{x}_3) \\ C_{21} &= \dot{x}_1 l_1 l_3 m_3 s_{23} + \dot{x}_1 l_1 l_2 m_2 s_2 + \dot{x}_1 l_1 l_2 m_3 s_2 - \dot{x}_3 l_2 l_3 m_3 s_3 \\ C_{22} &= -\dot{x}_3 l_2 l_3 m_3 s_3 \quad C_{23} = -l_2 l_3 m_3 s_3 (\dot{x}_1 + \dot{x}_2 + \dot{x}_3) \\ C_{31} &= l_3 m_3 (\dot{x}_1 l_2 s_3 + \dot{x}_2 l_2 s_3 + \dot{x}_1 l_1 s_{23}) \quad C_{32} = l_2 l_3 m_3 s_3 (\dot{x}_1 + \dot{x}_2) \\ C_{33} &= 0 \\ G_1 &= g (l_1 m_1 c_1 + l_1 m_2 c_1 + l_1 m_3 c_1 + l_3 m_3 c_{123} + l_2 m_2 c_{12} + l_2 m_3 c_{12}) \\ G_2 &= g (l_3 m_3 c_{123} + l_2 m_2 c_{12} + l_2 m_3 c_{12}) \\ G_3 &= g l_3 m_3 c_{123} \end{aligned}$$

## REFERENCE

- [1] D. T. Tran, M. Jin, and K. K. Ahn, "Nonlinear Extended State Observer Based on Output Feedback Control for a Manipulator With Time-Varying Output Constraints and External Disturbance," *IEEE Access*, vol. 7, pp. 156860-156870, 2019.
- [2] J. J. Craig, *Introduction to robotics: mechanics and control*. Pearson Prentice Hall Upper Saddle River, 2005.
- [3] T. D. Thien, D. X. Ba, and K. K. Ahn, "Adaptive Backstepping Sliding Mode Control for Equilibrium Position Tracking of an Electrohydraulic Elastic Manipulator," *IEEE Transactions on Industrial Electronics*, pp. 1-1, 2019.
- [4] D.-T. Tran, H.-V.-A. Truong, and K. K. Ahn, "Adaptive Backstepping Sliding Mode Control Based RBFNN for a Hydraulic Manipulator Including Actuator Dynamics," *Applied Sciences*, vol. 9, no. 6, p. 1265, 2019.
- [5] T. X. Dinh, T. D. Thien, T. H. V. Anh, and K. K. Ahn, "Disturbance Observer Based Finite Time Trajectory Tracking Control for a 3 DOF Hydraulic Manipulator Including Actuator Dynamics," *IEEE Access*, vol. 6, pp. 36798-36809, 2018.
- [6] N. Nikdel, M. Badamchizadeh, V. Azimirad, and M. A. Nazari, "Fractional-Order Adaptive Backstepping Control of Robotic Manipulators in the Presence of Model Uncertainties and External Disturbances," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 10, pp. 6249-6256, 2016.
- [7] Q. Guo, Y. Zhang, B. G. Celler, and S. W. Su, "Neural Adaptive Backstepping Control of a Robotic Manipulator With Prescribed Performance Constraint," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3572-3583, 2019.
- [8] T. J. J. van den Boom, F. o. M. E. TU Delft, and M. Technology, *Model Predictive Control: SC4060*. TU Delft, 2013.
- [9] K. P. Tee, S. S. Ge, and E. H. Tay, "Barrier Lyapunov Functions for the control of output-constrained nonlinear systems," *Automatica*, vol. 45, no. 4, pp. 918-927, 2009/04/01/ 2009.
- [10] W. Meng, Q. Yang, and Y. Sun, "Adaptive Neural Control of Nonlinear MIMO Systems With Time-Varying Output Constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1074-1085, 2015.
- [11] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789-814, 2000/06/01/ 2000.
- [12] K. P. Tee, B. Ren, and S. S. Ge, "Control of nonlinear systems with time-varying output constraints," *Automatica*, vol. 47, no. 11, pp. 2511-2516, 2011/11/01/ 2011.
- [13] C. P. Bechlioulis and G. A. Rovithakis, "Robust Adaptive Control of Feedback Linearizable MIMO Nonlinear Systems With Prescribed Performance," *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2090-2099, 2008.
- [14] J. Xia, J. Zhang, J. Feng, Z. Wang, and G. Zhuang, "Command Filter-Based Adaptive Fuzzy Control for Nonlinear Systems With Unknown Control Directions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1-9, 2019.
- [15] R. D. Nussbaum, "Some remarks on a conjecture in parameter adaptive control," *Systems & Control Letters*, vol. 3, no. 5, pp. 243-246, 1983/11/01/ 1983.
- [16] S. S. Ge, H. Fan, and L. Tong Heng, "Adaptive neural control of nonlinear time-delay systems with unknown virtual control coefficients," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 499-516, 2004.
- [17] Y. Wu, R. Huang, X. Li, and S. Liu, "Adaptive neural network control of uncertain robotic manipulators with external disturbance and time-varying output constraints," *Neurocomputing*, vol. 323, pp. 108-116, 2019/01/05/ 2019.

# Determine the percentage of Recovering FCC Spent Catalysts as Mineral Filler in the Asphaltic Concrete Mixture

Anh Thang Le  
Faculty of Civil Engineering  
University of Technical and Education  
Vietnam  
thangla@hcmute.edu.vn

Nguyen Manh Tuan  
Faculty of Civil Engineering  
Ho Chi Minh City University of Technology  
Vietnam

**Abstract**—Recovering fluid catalytic cracking (FCC) is an industrial waste produced from the cracking of petroleum in oil refineries. In Vietnam, it is discharged a lot from the Dung-Quoc and Nghi-Son oil refineries. Besides, asphalt concrete mixture consisting of aggregate, filler, and asphalt binder, is requested with a significant volume for the development of Vietnam. The natural filler usually is the limestone powder that causes the cost of asphalt concrete mixtures increasing. Finding the resource replacing for the limestone powder could decrease the cost of asphalt concrete. In this experimental study, the spent catalyst replaced a part of filler in asphalt concrete mixtures was performed. The study explored the possibility of using this material in road construction in the form of filler material in asphalt concrete mixtures. The optimum percentage of spent fluid catalytic cracking catalyst (SFCC) replacing natural filler to the point that the characteristics of the mixture not be degraded was founded based on the Marshall methods. As a result, mixtures that used 5% SFCC of the weight of hot mix asphalt (HMA) had the best Marshall characteristics.

**Keywords**— Recovering FCC spent catalyst, Asphalt mixture, Asphalt concrete filler

## I. INTRODUCTION

In the petrochemical industry, catalysts play an important role, especially the FCC (Fluid Catalytic Cracking). FCC catalysts have main components such as  $\text{SiO}_2$ ,  $\text{Al}_2\text{O}_3$ ,  $\text{La}_2\text{O}_3$ ,  $\text{TiO}_2$ ,  $\text{CaO}$ ,  $\text{Fe}_2\text{O}_3$ ,  $\text{NiO}$ ,  $\text{V}_2\text{O}_5$ ,  $\text{Na}_2\text{O}$ ,  $\text{MgO}$ ,  $\text{P}_2\text{O}_5$ ,  $\text{CeO}_2$  used from the early 1960s. FCC contributes to improving the efficiency of valuable products such as gasoline [1].

Dung-Quoc oil refineries discharge about 20 tons FCC catalysts. FCC catalyst is steadily increasing over the years. Meanwhile, Dung-Quoc Oil Refinery currently has no plan to recover the FCC catalysts. Compared with the spent fluid catalytic cracking catalyst (SFCC) in the world, the SFCC of Dung-Quoc oil refinery has a particular feature that contains very high Fe content [1].

The fluid catalytic cracking catalyst (FCC) technology utilized more and more due to the demand for fuel in particular, and energy, in general, is increasing. So the emitted amount of FCC catalysts is enormous. The problem of FCC catalytic disposal becomes essential. In similar to other hazardous waste, FCC catalysts are treated by a simple method which is landfilled and partly as construction materials.

SFCC was utilized in the concrete by many study works. For example, De Lomas (2007) [2] presents the compressive strength improving by blended mortars containing spent fluid

catalytic cracking catalyst (SFCC). SFCC was known that causes hydration heat increasing and producing pozzolanic activity. The experiment results show the hydration heat increasing in the case of mortar with 10% SFCC. The compressive strength of mortar with SFCC is higher than the reference mortar due to the pozzolanic effect. Due to pozzolanic characteristics, SFCC has been studied that causing the excellent performance of concrete mixtures using SFCC as cement replacement. Incorporating SFCC into Portland cement mortars and concrete helps to improve mechanical properties and durability. Hydration products in cement added SFCC were very similar to those produced in cement with a metakaolin system. The mechanical behavior of mortars added SFCC was similar to or even more superior than ones having added metakaolin [3]. In Oman, Torres Castellanos (2010) explored the possibility of using SFCC in asphalt concrete mixtures in the form of aggregate or filler material replacement. The experiment results show good potential for using spent catalyst in road applications [4].

Asphalt concrete is the preferred material in the pavement surface of Vietnam and the world. The asphalt concrete pavement surface has the advantage of driving comfort, durability, and water resistance [5, 6].

Asphalt concrete mixture is a combination of coarse particles, fine particles, mineral filler, and asphalt binder. The coarse and fine particles take approximately 90% of the weight of HMA. Filler in asphalt mixtures takes approximately 3% to 7% of the weight of HMA. The compacted aggregate acts as the structural skeleton of the asphalt concrete mixture and the asphalt binder as the glue of the asphalt concrete mixture. Properties of aggregate affect directly on the performance of asphalt pavements [7].

Flexible pavements are designed so that pavements have a long economic life. The design period of flexible pavement is approximately 20 years. During the design period, the pavement surface must low maintenance/repair costs, low-cost construction and repair time, compatibility with the environment. Furthermore, flexible pavements are rather high-cost structures. Thus the materials to be used for their constructions must be appropriately designed. The asphaltic concrete materials to be used in road construction have the primary role of the contribution to the economic achievement of the project. The waste materials utilized in the asphalt concrete mix need quality control procedures for the adaption of the HMA specified requirement. Waste materials are aimed to increase the performance and lifetime of roads [8].

The particle shape, texture, and mineral filler content in the mix are effect in permanent deformation of the flexible pavement [9]. For exploring the effect of different mineral filler rates, Tayebali, A.A. (1998) [9] conducted a test on the asphalt concrete mixtures containing 4, 6, 8, and 12% mineral filler designed by using the Marshall procedure. Their experiment results indicate that asphalt concrete mixtures containing 100% crushed granite replacing for natural sand lowered the accumulation of permanent strain than a blend of crushed granite and natural sand. Besides, an increase in the mineral filler content of a mixture could decrease the accumulated permanent strain of the mixture while increasing the mixture shear resilient modulus. Many studies explored the effect of filler in the asphalt concrete mixtures by changing the filler rate and gradually replacing filler by percentage.

Although many studies were conducted about using SFCC in many work areas, there are not many studies about the use of SFCC replacing a part of mineral filler in the asphalt concrete. In this study, the usability of SFCC in hot mix asphalt concrete was investigated based on Marshall mixtures design.

## II. MATERIALS

### A. Aggregates

The same type of crushed limestone aggregates was used in all asphalt mixtures in the study. Aggregate material was tested to obtain the mechanical characteristics based on Vietnamese Standards. The aggregate characteristics are given in Table 1.

In the study, aggregate grading curves of asphalt mixtures were chosen from Vietnamese Roadway Construction Specifications. Sieve analyses were carried out, and the grading curve for the aggregate used in the study was shown in Fig. 2. The upper and lower boundary curves of aggregate for asphalt concrete are the required boundary curves corresponding to the nominal maximum aggregate size of 12.5mm, according to TCVN 8819: 2011 [10].

TABLE I. PROPERTIES OF AGGREGATE

Sieve diameters	Properties	Standard	Aggregate
12.5-4.75 mm	Bulk specific gravity ( $\text{g}/\text{cm}^3$ )	ASTM C 128-88	2.684
	Apparent specific gravity ( $\text{g}/\text{cm}^3$ )		2.725
	Compressive strength (MPa)		$\geq 100$
	Abrasion loss (%) (Los Angeles)	ASTM C 131	14,64
	Flat and Elongated (%)	ASTM D4791	10,73
	Content of dust, mud, clay (based on the total aggregate weight) (%)		$< 2$
4.75-0.075 mm	Bulk specific gravity ( $\text{g}/\text{cm}^3$ )	ASTM C 127-88	2.647
	Apparent specific gravity ( $\text{g}/\text{cm}^3$ )		2.694
Filler	Specific gravity ( $\text{g}/\text{cm}^3$ )		2.739

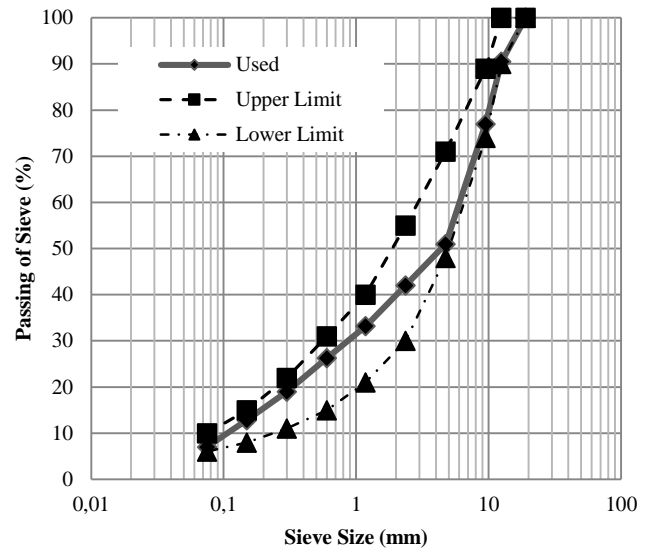


Fig. 1. Grading curves of aggregate.

### B. Asphalt binder

The asphalt binder used in the study is a conventional asphalt in Vietnam with a 60–70 penetration. Bitumen has the mechanical and physical criteria that meet the Vietnamese standard TCVN 7493: 2005 [11]. The physical characteristics of this bitumen were given in Table 2.

TABLE II. PHYSICAL CHARACTERISTICS OF THE BITUMEN 60/70.

Test name	Units	Values	TCVN 7493:2005
Penetration (25°C)	0.1mm	60.7	Min. 60, Max. 70
Ductility (25°C, 5cm/min)	cm	$>100$	Min. 100
Softening point	°C	49.3	Min. 46
Fire point	°C	318	Min. 232
Bitumen loss on the heating test (a temperature of 163°C for 5 hours)	%	0.14	Max. 0.8
The ratio of penetration compared to the original (after heating 5 hours at 163°C)	%	79.46	Min. 75
Solubility test (using trichloroethylene)	%	99.83	Min.99
Specific gravity 25°C.	$\text{g}/\text{cm}^3$	1,037	1-1,05
Kinematic viscosity (at 60°C)	Pa.s	276	Min.180
The wax content	%	1.52	Max. 2.2
The adhesion between stones and bitumen	level	3	Min. level 3

### C. Filler

a) *Spent fluid catalytic cracking*: The morphology of FCC is explored through the SEM images (Fig. 2a). It was observed that the original FCC was very smooth; most particles were not larger than 20  $\mu\text{m}$ . The SFCC had been broken compared to the original FCC catalyst (the average size usually is about 60-70  $\mu\text{m}$  [1, 12]). The tiny particle size of SFCC could favorable for filler or increasing strength asphalt concrete mix as a cement additive. The chemical composition of SFCC was determined by the EDX method. The oxides in the SFCC sample are mainly  $\text{Al}_2\text{O}_3$  and  $\text{SiO}_2$ , with the percentage masses of about 50.74% and 43.18% [1], respectively.

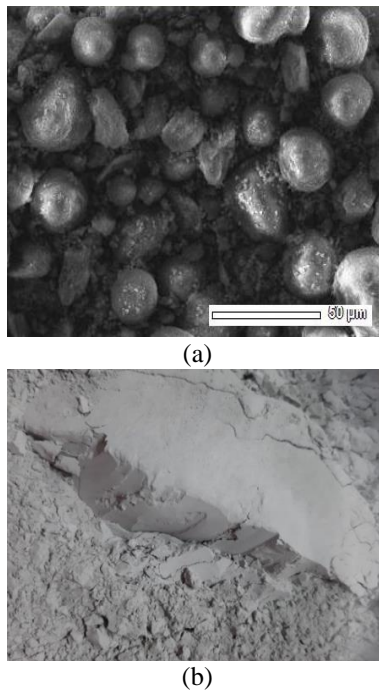


Fig. 2. Spent fluid catalytic cracking: (a) SEM [1], (b) SFCC Filler.

*b) Limestone:* The limestone (LS) filler used in this study was INSEE-Bitu-Fill. It is produced with guarantees of stable quality. The INSEE-Bitu-Fill was known as the filler to ensure consistent and quality asphalt performance for infrastructure in Vietnam. It has a high surface area. The specific surface area is 3,000 cm<sup>2</sup>/g.

#### D. Marshall method

Samples were prepared with a diameter of 101.6mm and an average height of 63.5mm by the Marshall mix design method. Each bitumen content ratio of 4.5%; 5%; 5.5%; 6%; 6.5% have three samples. SFCC was added directly to the aggregate and limestone filler after it was placed in the oven at 90°C in 24 hours for eliminating the moisture. Marshall Test was performed according to TCVN 8860-1: 2011 [9].

#### E. Flow chart of the study

Several asphalt concrete samples of various LS filler proportions (4%, 5%, 6%, and 7%) were produced and tested by the Marshall method. As a result, the asphalt concrete mix with 5% LS filler was selected for further study. The chosen asphalt concrete mix has given high stability with the lowest asphalt binder content.

A flowchart described the approach of the study was presented in Fig. 3. The figure shows that asphalt concrete samples were prepared for the selected filler proportion of 5% with five different bitumen content (4.5%, 5%, 5.5%, 6%, and 6.5%). These samples were tested with the Marshall mix design method for determining the amount of optimum bitumen. Marshall Stability, flow value, void volume values (Vh), void filled with bitumen (VFA), and voids in mineral aggregate (VMA) values were determined.

Marshall Test was performed on the asphalt concrete mixture with 0%, 25%, 50%, 75%, and 100% SFCC partly replacing for LS. As a result, 5% SFCC by weight of the mixture was the recommended optimum percentage of SFCC.

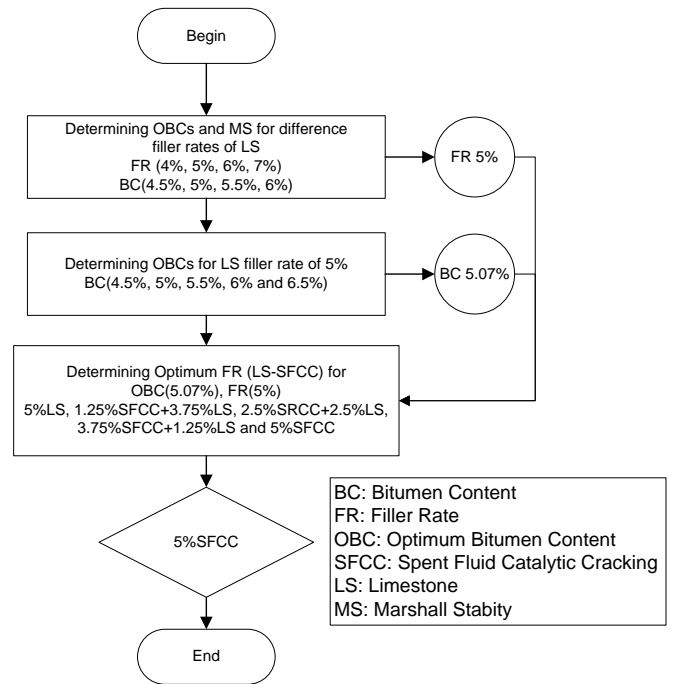


Fig. 3. Flow chart of the study.

### III. EXPERIMENTAL RESULTS AND DISCUSSION

#### A. Determining optimum filler rates of limestone

Fig. 4. shows the variation of OBC following the limestone filler rates. As can be seen from the figure, the increasing the filler rates was increasing the OBC and increasing the cost for asphalt concrete mix. In terms of effective asphalt content, the film thickness covering an aggregate particle will be decreased if the aggregate is fine gradation. Thus, more asphalt content is required for increasing the percentage of filler. In this study, bitumen content increased, whereas LS mineral filler in mixtures gave reductions less than 5%, as shown in Fig. 5. Voids filled with bitumen increased leads the bitumen content increased. The minimum OBC value was found with the asphalt concrete mix having a 5% filler rate (FR).

Besides, Marshall Stability was moderate-high (Fig. 5), and other Marshall characteristics met the requirement of TCVN 8860-1: 2011 [9]. Thus, the 5% filler rate was used in the study.

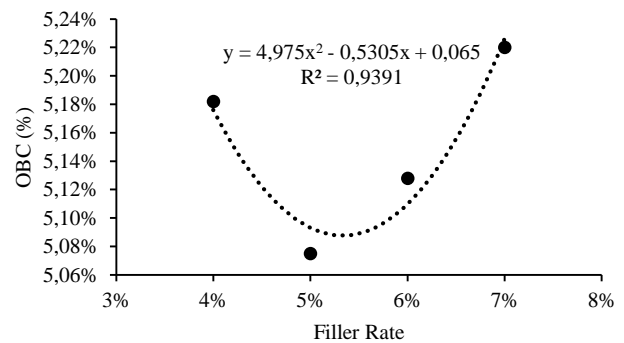


Fig. 4. Change in optimum binder content for different LS filler rates.



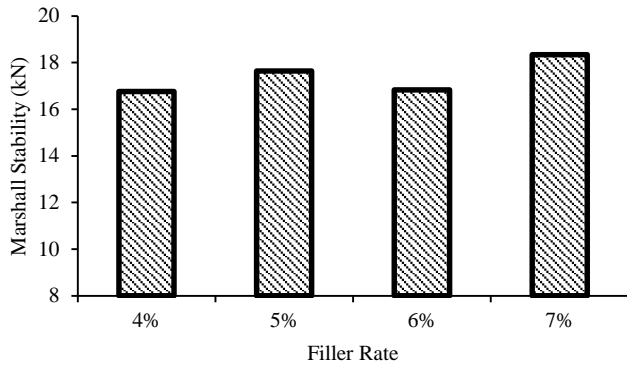


Fig. 5. Change in Marshall Stability for different LS filler rates.

### B. Investigating asphalt concrete mixture with the limestone filler rate of 5%

The Marshall Test was utilized to checking for the optimum percentage of SFCC in an asphalt concrete mix because the asphalt concrete properties obtained from the Marshall test, which could be determined following the Marshall Test procedure, contented the information relating to the asphalt concrete performance.

For any given asphalt and aggregate mixture, the durability is enhanced if the adequate film thickness is attained. Adequate film thickness could avoid excessive asphalt bleeding or flushing. The adequate Voids in Mineral Aggregate (VMA) have the relationship to the adequate film thickness [13].

The Marshall Flow value reflects the plasticity and flexibility properties of asphalt mixtures. Marshall flow has a linear inverse relationship with internal friction [14].

Fig. 6c shows the relationship between flow and bitumen content. The Marshall flow of the mixture increased with asphalt bitumen content increased, as a result of the experiment showed a nonlinear relationship.

The optimum bitumen content was determined based on the specific asphalt bitumen contents estimated for meeting the requirement of the specification. It was evaluated as the average value of the following five bitumen contents taken from the graphs in Fig. 6.

- Bitumen content of the maximum stability value in Fig. 6b (i.e. 4.5%).
- Bitumen content of the maximum bulk specific gravity in Fig. 6a (i.e. 5.5%).
- Bitumen content corresponding to the percentage air voids closing to a percentage of 4% (i.e. 4.74%, Fig. 6d).
- Bitumen content corresponding to the median of percentage voids filled with bitumen (i. e. 5.31%, Fig. 6f).
- Bitumen content corresponding to the median of flow (i. e. 5.32%, Fig. 6c).

The optimum bitumen content:

$$\frac{4.5 + 5.5 + 4.75 + 5.31 + 5.32}{5} = 5.07\%$$

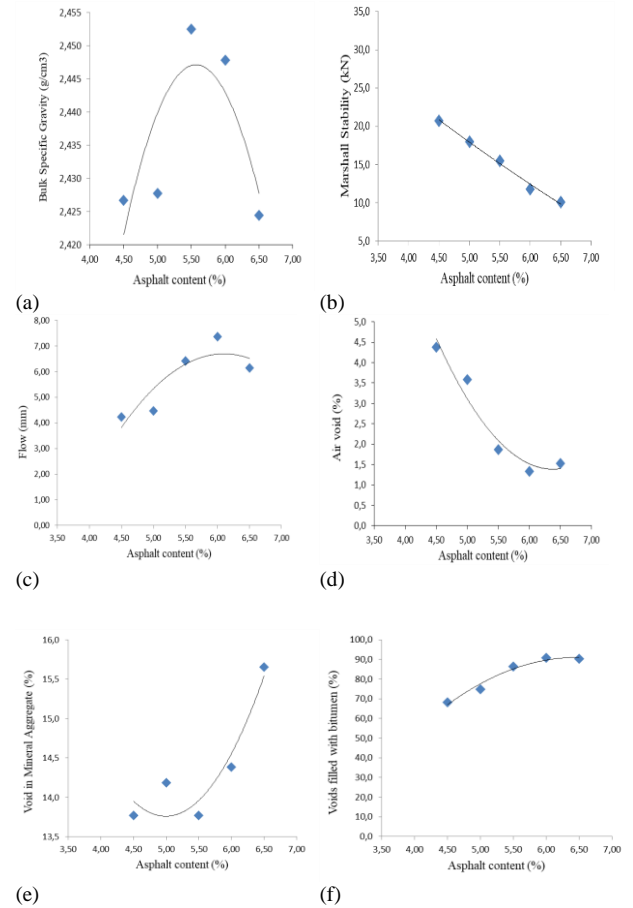


Fig. 6. Comparative charts of Marshall test.

### C. Determining optimum SFCC partly replacement for limestone

The Marshall characteristics obtained for all mixtures containing 5% filler rate, in which SFCC partly replaced for LS, were given in Figs. 7–12. Marshall Stability values of all samples were higher than the specification limit of 8kN [10] (Fig. 7). As seen from Marshall flow values, obtained bitumen content, and the SFCC rate above 2.5%, the flow values were in the specification range (2mm-4 mm) (Fig. 8).

Based on TCVN 8819:2011 [10] and MS-2 [13], VMA must be higher than 14%, the percentage air voids (Vh) should be reached to a percentage of 4%, and VFA is in a range of 65% to 75%. As seen from Figs. 10, 11, and 12, Vh, VMA, and VFA values of the samples prepared with 2.5% SFCC and 2.5% LS were satisfied with the specification limits.

Maximum Marshall Stability value (21.99 kN) obtained from 5.07% bitumen content and 5% filler rate (3.75% SFCC + 1.25% LS). It was seen an increase of Marshall Stability for the samples with a high rate of SFCC (Fig. 7). 15.5 kN, which is the lowest Marshall Stability value, has measured on samples prepared with 5% LS filler, while 21.72 kN, which is the Marshall Stability value, has measured on samples prepared with 5% SFCC. Based on the Marshall Stability value, the SFCC rate of 5% is recommended as the best filler rate of the asphalt concrete mix in the study.

When considering flow charts, the lowest flow value was obtained at the point of 3.0 mm, corresponding to 5% SFCC, and the highest flow value was obtained at the point of 6.42

mm, corresponding to 5% LS filler. When comparing mixtures full with LS, flow value has a 53.50% decrease for samples prepared with 5% SFCC, and 37.69% decrease for samples prepared with 2.5% SFCC and 2.5% LS.

When considering bulk specific gravity (SG) charts with the increase in SFCC proportion in filler rate, there is a decrease in SG value. While SG value is 2.453 kg/cm<sup>3</sup> for asphalt samples prepared with 5% LS, with the increase in substitution proportion, there is a decrease in SG values and min value of SG, which is 2.393 kg/cm<sup>3</sup> was obtained from asphalt samples prepared with 5% SFCC. There is a maximum 2.4% decrease of SG value according to the reference samples.

When considering VMA charts with the increase in SFCC proportion in filler rate, there is an increase in VMA value. While the VMA value is 13.62% for asphalt samples prepared with 5% LS and 0% SFCC, with the increase in substitution proportion, there is an increase in VMA values and the max value of VMA, which is 15.71% was obtained from asphalt samples prepared with 5% SFCC and 0% LS. There is a maximum 15.3% increase of VMA value according to the reference samples.

When considering air void values with the increase in SFCC proportion in filler rate, there is an increase in V<sub>h</sub> values. While the V<sub>a</sub> value is 2.48% for samples prepared with 5% LS, with the increase in substitution proportion, there is an increase in V<sub>h</sub> values and max value of V<sub>h</sub>, which is 4.84% for samples prepared with 5% SFCC. There is a maximum 91.71% increase of V<sub>h</sub> value according to the reference samples.

When considering VFA charts with the increase in SFCC proportion in filler rate, there is a decrease in VFA value. While VFA value is 81.76% for asphalt samples prepared with 5% LS, with the increase in substitution proportion, there is a decrease in VFA values, and the min value of VFA, which is 69.2% was obtained from asphalt samples prepared with 5% SFCC. There is a maximum 15.36% decrease of VFA value according to the reference samples.

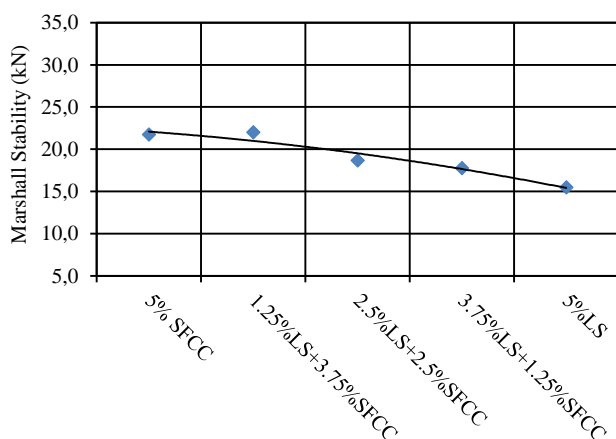


Fig. 7. Change in Marshall Stability for different SFCC contents.

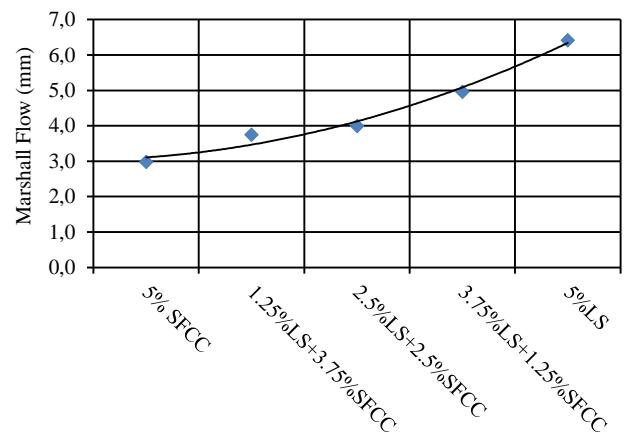


Fig. 8. Change in Marshall Flow for different SFCC contents.

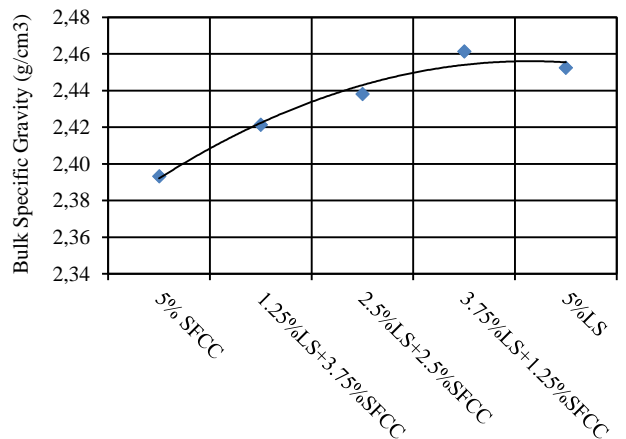


Fig. 9. Change in Bulk Specific Gravity for different SFCC contents.

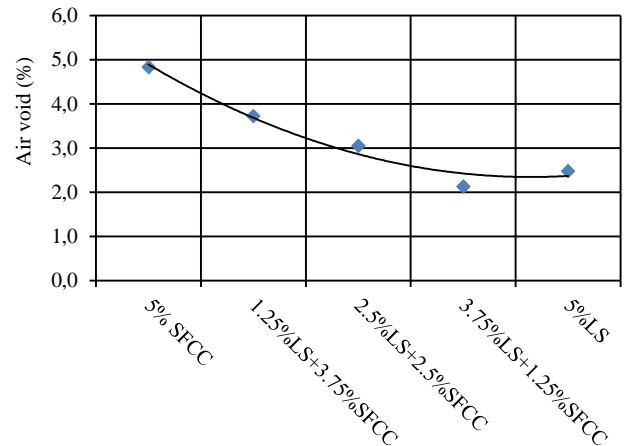


Fig. 10. Change in Air void for different SFCC contents.

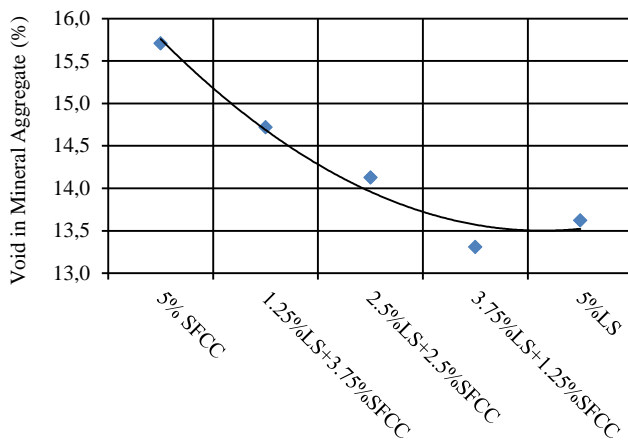


Fig. 11. Change in VMA for different SFCC contents.

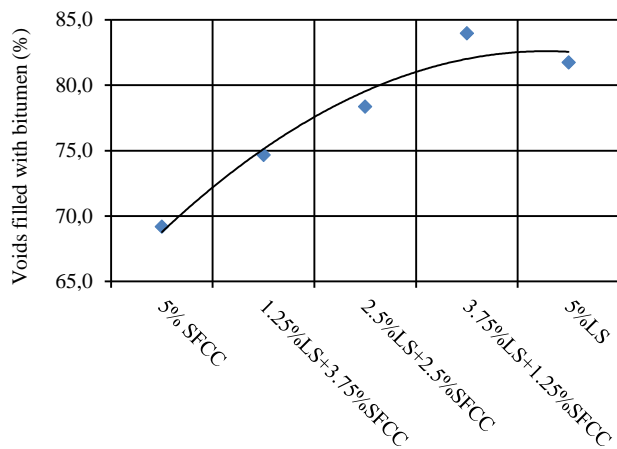


Fig. 12. Change in VFA for different SFCC contents.

#### IV. MATERIALS

In the study first, four different proportions of filler rate (4%, 5%, 6%, and 7%) were chosen based on the maximum and minimum filler rate of the conventional asphalt mixture in Vietnam. Optimum bitumen contents were determined for every filler rate with the test based on the Marshall Test procedure given in TCVN 8819:2011. As a result, the minimum optimum bitumen contents (OBC) and moderate-high Marshall Stability values were determined as 5.07% for a 5% filler rate.

In the first section, tests were repeated on samples prepared with determined OBC, and comparative charts were generated. Results have shown that the 5.07% OBC prepared with 5% filler rate has given the most successful result.

In the second section of the study, SFCC obtained from Dung-Quoc oil refineries was changed with a determined 5% mineral filler in the proportion of 25%, 50%, 75%, and 100%. The optimum SFCC substitution ratio was determined by comparing the results.

Test results have shown that mixtures that used 5% SFCC of the weight of HMA had the best Marshall stability, and other characteristics met the specification requirement. A recommended was given that SFCC could replace 100% for the conventional mineral filler in the asphalt concrete mixture of Vietnam.

#### ACKNOWLEDGMENT

This work belongs to the project grant No: B2019-SPK-01. funded by Ministry of Education and Training, and hosted by Ho Chi Minh City University of Technology and Education, Vietnam.

#### REFERENCES

- [1] T. T. Dang, P.H.N., "Research on using the FCC catalyst for the cracking reaction of WAX dregs from the pyrolysis process of waste polypropylene plastic," *Journal of MOIT*, 2014.
- [2] De Lomas, M.G., M.S. De Rojas, and M. Frías, "Pozzolanic reaction of a spent fluid catalytic cracking catalyst in FCC-cement mortars," *Journal of Thermal Analysis and Calorimetry*, vol. 90(2), pp. 443-447, 2007.
- [3] Torres Castellanos, N. and J. Torres Agredo, "Using spent fluid catalytic cracking (FCC) catalyst as pozzolanic addition a review," *Ingeniería e investigación*, vol. 30(2), pp. 35-42, 2010.
- [4] Alshamsi, K., et al., "Utilizing waste spent catalyst in asphalt mixtures," *Procedia-Social and Behavioral Sciences*, vol. 53, pp. 326-334, 2012.
- [5] Uzun, İ. and S. Terzi, "Evaluation of andesite waste as mineral filler in asphaltic concrete mixture," *Construction and Building Materials*, vol. 31, pp. 284-288, 2012.
- [6] Yilmaz, M., B.V. K k, and N. Kulo lu, "Effects of using asphaltite as filler on mechanical properties of hot mix asphalt," *Construction and Building Materials*, vol. 25(11), pp. 4279-4286, 2011.
- [7] Ahmedzade, P. and B. Sengoz, "Evaluation of steel slag coarse aggregate in hot mix asphalt concrete," *Journal of hazardous materials*, vol. 165(1-3), pp. 300-305, 2009.
- [8] Serin, S., et al., "Investigation of usability of steel fibers in asphalt concrete mixtures," *Construction and Building Materials*, vol. 36, pp. 238-244, 2012.
- [9] Tayebali, A.A., G.A. Malpass, and N. Paul Khosla, "Effect of mineral filler type and amount on design and performance of asphalt concrete mixtures," *Transportation research record*, vol. 1609(1), pp. 36-43, 1998.
- [10] MOST, TCVN 8819:2011 in Specification for Construction of Hot Mix Asphalt Concrete Pavement and Acceptance, 2011.
- [11] MOST, TCVN 7493:2005 in Bitumen Specifications, 2005.
- [12] Scherzer, J., Octane-enhancing, "Zeolitic FCC catalysts: scientific and technical aspects," *Catalysis Reviews—Science and Engineering*, vol. 31(3), pp. 215-354, 1989.
- [13] Institute, A., MS-2 asphalt mix design methods. USA: Lexington Kentucky, 2014.
- [14] Umar, F. and E. A ar, Pavement structure. İstanbul Technical University Civil Engineering Faculty Press, 1991.

# The Effect Of Capital Structure On Profitability: An Empirical Analysis Of Vietnamese Listed Banks

Tran Thuy Ai Phuong

Department of Accounting and Finance  
HCMC University of Technology and  
Education

Ho Chi Minh City, Vietnam  
phuongtta@hcmute.edu.vn

Nguyen Thi Anh Van

Department of Industrial Management  
HCMC University of Technology and  
Education

Ho Chi Minh City, Vietnam  
anhvan@hcmute.edu.vn

Nguyen Thi Hoang Anh

Department of Accounting and Finance  
HCMC University of Technology and  
Education

Ho Chi Minh City, Vietnam  
anhnh@hcmute.edu.vn

**Abstract**— This paper aims at investigate the impact of capital structure on profitability of seventeen listed banks in Vietnam over a period from 2015 to 2018. The profitability is measured by Return on Total Assets (ROA) while capital structure is measured by Debt to Total Assets (DA), firm's size, growth opportunity, and tangibility. The data has been analyzed by using descriptive statistics, correlation, and panel data regression models. The findings show that there is a strong negative impact of capital structure on firm's profitability. These findings offer useful insights for the banks and lending institutions based on empirical evidence.

**Keywords**— Capital Structure, Profitability, Profit, Debt, Total Assets, Return on Total Assets.

## I. INTRODUCTION

Capital structure is one of the most puzzling issues in corporate finance literatures [1]. This term is defined as a combination of debt and equity in a company. Deciding on an optimal capital structure is one of many strategic decisions made by corporate managers as it helps to reduce the cost of capital and directly affects the firm's profitability [2].

The above issue has received the attention of many scholars around the world. However, the current research results are hitherto inconsistent. Many studies found that the capital structure has a negative impact on the profitability of the firm [2] [3] [4] [5]. On the other hand, the findings of Abor (2005) [6]; Gill, Nahum, and Neil (2011) [7]; Suleiman M. Abbadi and Nour Abu Rub (2012) [8]; Singh and Bagga (2019) show that the company's debt-to-assets ratio and the profitability are positively correlated [9].

In the wake of integration and globalization, Vietnam is transforming itself in the renovation process, the economy is operating under the market mechanism, with the regulation and macro management of the government. To survive and rapidly grow, enterprises and commercial banks must ensure efficient operations, hence, the profitability is a top concern. Up to now, there has been very little research on the effect of capital structure on the profitability of Vietnamese listed banks.

Studying the effect of capital structure on the profitability of Vietnamese listed banks is important both academically and practically. Academically, the study clarifies the relationship between capital structure and the profitability of banks. Practically, this study provides important implications for banks in governance, assists government agencies manage the activities of commercial banks as well as helps investors in

analyzing and giving investment decisions. Therefore, the research issue: "The Effect of Capital Structure on Profitability: An Empirical Analysis of Vietnamese Listed Banks" need to be implemented.

## II. LITERATURE REVIEW

### A. International studies

The study of Abor (2005) was conducted with a sample size of 20 enterprises listed on the Ghana stock exchange (GSE) [6]. The purpose is to understand the relationship between capital structure and profitability. The results show a positive relationship between short-term debt and Return on Equity (ROE) and a negative relationship between long-term debt and ROE. In addition, this study also shows that total debt and ROE have a positive relationship too.

Abor's study paves the way for a series of later studies. Gill, Nahum, and Neil have conducted a research with a sample of 272 American firms listed on New York Stock Exchange. The correlations and regression analyses were used to estimate the relationship between capital structure and profitability. Empirical results show positive impacts of short-term debt to total assets (SDA), long-term debt to total assets (LDA) and total debt to total assets (DA) on ROE in the manufacturing industry. The paper also indicates that SDA and DA have positive effects on the profitability in the service industry [7].

The research of Mohammad Fawzi Shubita and Jaafer Maroof Alsawalhah also seeks to extend Abor's and Gill's findings regarding the effect of capital structure on profitability. Data sources used include financial data of 39 industrial companies listed on the Amman Stock Exchange over a 6-year period (2004 to 2009). The results show a significantly negative relation between debt and profitability. The findings also show that the profitability increases with control variables [4].

Suleiman M. Abbadi and Nour Abu-Rub (2012) published their research based on data collected from 28 Palestinian financial institutions during 2006-2010. The ordinary least squares (OLS) and Multiple Linear Regression (MLR) were employed to investigate the effect of capital structure on the performance of these institutions. They found strong correlation between return on assets and efficiency; and total deposit to total assets and efficiency [8].

Mahfuzah Salim and Raj Yadav (2012) conducted research with a panel data of 237 Malaysian companies listed

on Bursa Malaysia Stock Exchange during 1995-2011. The findings show that ROA, ROE, EPS are inversely correlated with short-term debt, long-term debt, and total debt. Tobin's Q ratio has a significant positive correlation with short-term debt and long-term debt [3].

The research of Singh and Bagga was to investigate the impact of capital structure on the profitability of Nifty 50 companies listed on National Stock Exchange of India during 2008 – 2017. In this study, Singh and Bagga created 4 different regression models to study the individual effect of total debt and total equity ratios on profitability, including ROA and ROE. All of these models have been tested with pooled OLS, fixed effects (FEM), and random effects (REM). The conclusion is that capital structure has a significant positive effect on firm's profitability [9].

### B. Domestic studies

The study of Tran Thuy Minh Chau (2018) was implemented with a sample size of 566 listed companies on Hochiminh Stock Exchange (HOSE) and Hanoi Stock Exchange (HNX) in 10-year period (2005- 2014) (not including financial institutions) to test the influence of capital structure on the profitability of these companies. Research results show that capital structure (including SDA, LDA, and DA) has negative relationship with ROA and ROE. The findings are contrary to the theory of Modigliani and Miller (1963) that the company will take advantage of the tax shield when using more debt [5] [10].

Tran Thi Bich Ngoc, Nguyen Viet Duc, and Pham Hoang Cam Huong investigated the effect of capital structure on the performance using a panel data sample of 130 joint stock companies in Thua Thien Hue province during 5-year period (2010-2014). The findings indicates that the capital structure has a significantly inversely impact on ROE, ROA and Earning per Share (EPS). Besides, it also shows that firm size affects positively on ROA and EPS. Finally, growth opportunities and asset structure were found to be inversely related to ROE and ROA [11].

The studies mentioned above show that there are many conflicting conclusions about the impact of capital structure on profitability. Therefore, testing the effect of capital structure on the profitability of Vietnamese listed banks is crucial. This study helps to assess whether the effect of capital structure on profitability in Vietnamese banking industry is consistent with other research results.

Research by Tran Thuy Minh Chau (2018) assessed the impact of capital structure on profitability of listed companies on Vietnamese stock exchange, including companies in many industries. Inheriting this research model, the authors conducted research on the banking industry only.

## III. MODEL AND METHODOLOGY

### A. Data and variables

In this research, the emphasis is on the impact of capital structure on the profitability of 17 commercial banks listed on HOSE and HNX for 4-year period (2015 - 2018). The selection of the study period is subject to the availability of required data. These 17 banks are listed in Table I.

TABLE I. LIST OF 17 BANKS

Stock code	Bank's name
ACB	Asia Commercial Joint Stock Bank
BAB	Bac A Commercial Joint Stock Bank
BID	Joint Stock Commercial Bank for Investment and Development of Vietnam
CTG	Vietnam Joint Stock Commercial Bank for Industry and Trade
EIB	Vietnam Export Import Commercial Joint Stock Bank
HDB	Housing Development Commercial Joint Stock Bank
KLB	Kien Long Commercial Joint Stock Bank
LPB	Lien Viet Post Joint Stock Commercial Bank
MBB	Military Commercial Joint Stock Bank
NVB	National Citizen Commercial Joint Stock Bank
SHB	Saigon Hanoi Commercial Joint Stock Bank
STB	Saigon Thuong Tin Commercial Joint Stock Bank
TCB	Vietnam Technological and Commercial Joint Stock Bank
TPB	Tien Phong Commercial Joint Stock Bank
VCB	Joint Stock Commercial Bank for Foreign Trade of Vietnam
VIB	Vietnam International Commercial Joint Stock Bank
VPB	Vietnam Prosperity Joint-Stock Commercial Bank

Source: Authors' own collection

Table II below describes all variables used in this study.

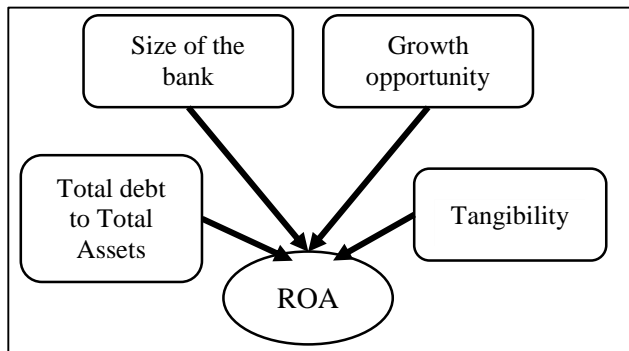
TABLE II. DEFINITION OF VARIABLES USED

Abbreviation	Variables	Definition	Authors
ROA	Return on Total Assets	Net income/ Average Total Assets	Abor (2005); Mahfuzah and Raj (2012); Tran Thi Bich Ngoc et al. (2017); Tran Thuy Minh Chau (2018); Singh and Bagga (2019)
DA	Total Debt to Total Assets	Total Debt/Total Assets	Abor (2005); Suleiman and Nour (2012); Mahfuzah and Raj (2012); Mohammad and Jaafer (2012); Tran Thuy Minh Chau (2018)
SIZE	Firm's size	Logarithm of total assets	Abor (2005); Mahfuzah and Raj (2012); Mohammad and Jaafer (2012); Tran Thi Bich Ngoc et al. (2017); Tran Thuy Minh Chau (2018)
GROWTH	Growth opportunity	$(\text{Net Sales}_n - \text{Net Sale}_{n-1}) / \text{Net Sale}_{n-1}$	Abor (2005); Mohammad and Jaafer (2012); Mahfuzah and Raj (2012); Tran Thuy Minh Chau (2018)
TANG	Tangibility	Fixed Assets/ Total Assets	Tran Thuy Minh Chau (2018)

Source: Authors' own collection



Referring to previous studies, the authors used ROA as the profitability indicator [6] [7] [8]. DA is the independent variable, representing capital structure [6]. In addition, SIZE [6], GROWTH [6] [7] [12] and TANG [12] are also included in the model as control variables. The research model was built as follows:



Source: Authors' own collection

Figure 1. Research model

Asymmetric information occurs when one of the parties to an economic transaction has more vital knowledge than the others. According to this theory, the managers have more information about the companies they are managing than outside investors or creditors. Therefore, they see the adjustment of capital structure as a signal of the information being held by managers [13]. Pettit and Singer (1985) believe that smaller businesses will have higher asymmetric information because the quality of their financial statements is not as high as that of large enterprises. On the other hand, the cost of capital mobilization (especially shares issue) of small businesses is often higher than that of larger enterprises. In addition, the issuance of new shares to raise capital will dilute the ownership of current shareholders. This is especially serious for small and medium-sized enterprises where current shareholders have to face the loss of control or even be annexed or merged [14]. Therefore, based on Pecking Order theory [15], the following hypotheses were formulated for the research.

H1. Capital structure (measured by DA) has a negative effect on the bank's profitability.

H2. The bank's size (measured by SIZE) has a positive effect on the bank's profitability.

H3. The growth opportunity (measured by GROWTH) has a positive effect on the bank's profitability.

H4. The tangibility (measured by TANG) has a positive effect on the bank's profitability.

## B. Techniques used

### 1) Descriptive statistics

Descriptive statistics help describe and understand the nature of a particular data set by providing brief summaries of the sample. Descriptive statistics include measures of central tendency and measures of variability. Measures of central tendency consist of mean, median, and mode. Whereas, measures of variability include the standard deviation, variance, the minimum and maximum variables, the kurtosis and skewness.

### 2) Correlation Analysis

Correlation represents the relationship between two variables. The higher the correlation, the stronger the relationship between the two variables. In addition to showing the strength of associations, correlation also indicates the direction of associations between two variables.

### 3) Regression Analysis - Pooled OLS, Fixed Effects, Random Effects

The authors used this technique to examine the strength of the impact of independent variables (DA, SIZE, GROWTH, and TANG) on dependent variable (ROA). The research model in this paper is defined as follows:

Profitability =  $f$  (Debt to Total Assets, firm's size, growth opportunity, tangibility)

$$ROA_{it} = \beta_0 + \beta_1 DE_{it} + \beta_2 DA_{it} + \beta_3 SIZE_{it} + \beta_4 GROWTH_{it} + \beta_5 TANG_{it} + u_{it}$$

The authors used some methods to test the model, including pooled ordinary least-squares (OLS), fixed effects (FEM), and random effects (REM). The Hausman test was then carried out to select the most appropriate method. Next, multicollinearity, serial correlation, and heteroskedasticity test were performed to find the defects of the model. Finally, the model's defects (if any) will be overcome to ensure the reliability of regression results.

## IV. RESULTS AND ANALYSIS

### A. Descriptive statistics

Table III describes the variables used in this research. Mean of ROA is 0.9%. The average capital structure of these banks is 92.6%. The majority of these institutions' debts are borrowed from the State Bank of Vietnam, other credit institutions and depositors. In the banking sector, a relatively high DA ratio is commonplace. Banks carry higher debt burden because they own a substantial amount of fixed assets in the form of a branch network.

TABLE III. DESCRIPTIVE STATISTICS

Variables	Mean	Minimum	Maximum	Std. Deviation
ROA	0.009	0.0001	0.289	0.006
DA	0.926	0.839	0.962	0.023
SIZE	8.318	7.403	9.118	0.412
GROWTH	0.228	-0.212	0.924	0.191
TANG	0.005	0.0009	0.019	0.004

Source: Authors' own calculation

### B. Correlation matrix

The correlations between variables are presented in Table IV. The results show that there is a strong negative correlation (-45%) between DA and ROA.

Besides, the table shows that there is a possibility that multicollinearity occurs between DA and SIZE, GROWTH and TANG. However, the use of the correlation matrix is rarely used for multicollinearity, mainly using the VIF command to check. The authors will describe the use of this command later.

TABLE IV. CORRELATION MATRIX

	ROA	DA	SIZE	GROWTH	TANG
ROA	1.0000				
DA	-0.4528	1.0000			
SIZE	0.2380	0.3480	1.0000		
GROWTH	0.3863	0.1390	0.0231	1.0000	
TANG	-0.3231	-0.1956	-0.1669	-0.3109	1.0000

Source: Authors' own calculation

### C. Regression results

The authors performed panel data regression using 3 methods, namely Pooled OLS, FEM and REM. Then the Hausman tests show that FEM is more appropriate in explaining the effect in the model. Thus, the authors follow the results of FEM only and discuss in detail.

In Table V, it is easy to see that DA has a negative impact on ROA at 5 percent level of significance. In other words, an increase in total debt results in a decrease in return on assets. Similar results are reported by Mahfuzah Salim and Raj Yadav, 2012 [3]; Mohammad Fawzi Shubita and Jaafer Maroof Alsawalhah, 2012 [4]; Tran Thuy Minh Chau, 2018 [5]; Tran Thi Bich Ngoc, Nguyen Viet Duc, and Pham Hoang Cam Huong, 2017 [14]. For control variables, only SIZE variable has a positive effect on ROA, all the remaining variables (including GROWTH and TANG) have no effect on ROA at 5 percent level of significance.

TABLE V. RESULTS OF REGRESSION ANALYSIS

	OLS	FEM	REM
DA	-0.1848 (0.000)	-0.1679 (0.000)	-0.1848 (0.000)
SIZE	0.0083 (0.000)	0.0255 (0.000)	0.0083 (0.000)
GROWTH	0.0078 (0.001)	0.0037 (0.072)	0.0078 (0.001)
TANG	-0.4746 (0.003)	-0.6139 (0.062)	-0.4746 (0.003)
Hausman	0.0000		

Source: Authors' own calculation

### D. Multicollinearity, serial correlation, and heteroskedasticity test

After performing regression, the authors conducted multicollinearity test with VIF command. The results show that all VIFs are smaller than 2 (Table VI). Therefore, the multicollinearity did not occur.

TABLE VI. MULTICOLLINEARITY RESULT

Variable	VIF	1/VIF
DA	1.18	0.850959
SIZE	1.16	0.865427
GROWTH	1.12	0.893556
TANG	1.15	0.866810
Mean VIF	1.15	

Source: Authors' own calculation

Using the xtserial command to conduct Wooldridge test for the serial correlation in the model, the authors then got the result in Fig.2. This proves that the serial correlation exists.

```
Wooldridge test for autocorrelation in panel data
H0: no first order autocorrelation
F( 1, 16) = 22.923
Prob > F = 0.0002
```

Source: Authors' own calculation

Figure 2. Wooldridge test result

The study used the xttest3 command to test for heteroskedasticity. The result is as shown in Fig. 3. The conclusion is that the FEM method did not encounter the heteroskedasticity.

```
Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (17) = 288.82
Prob>chi2 = 0.0000
```

Source: Authors' own calculation

Figure 3. Modified Wald test result

### E. Generalized method of moments (GMM)

After checking the model's defects and fixing them, the regression results (Table VII) show that DA still has negative effect on the ROA. This is not in conflict with the regression results in section C.

TABLE VII. REGRESSION ANALYSIS RESULTS - GMM

	Co.ef	P value
DA	-0.1086	0.071
SIZE	0.0304	0.009
GROWTH	0.0044	0.131
TANG	-0.4538	0.571

Source: Authors' own calculation

## V. RESEARCH LIMITATIONS

Many new legal regulations in the banking industry, which will officially take effect from the end of 2019, will have different impacts on banks' business results. For example, the regulation of reducing deposit interest rates on short-term deposits and lending interest rates for some sectors (effective from November 2019) could have an immediate impact on profits of banks. Future studies should include variable POLICY, representing policy changes in Vietnam, into the model for a more comprehensive assessment.

## VI. CONCLUSION AND RESEARCH IMPLICATION

In this article, the authors analyze the effect of capital structure on the profitability of 17 listed banks in Vietnam from 2015 to 2018. Descriptive statistics shows that all these banks were using too much debt. Correlation analysis indicates that DA has a negative effect on ROA.

To select the appropriate method, the study used regression model with panel data with Pooled OLS, FEM, REM, and Hausman test to select the most appropriate one.

Multicollinearity, serial correlation, and heteroskedasticity test were also implemented to find the defects of the model. The results after fixing the defects show that the capital structure is inherently an important factor and has a negative impact on the profitability of the banks. The results of the fixed effect model show that an increase in total debt results in a decrease in return on assets. Similar results are reported by Mahfuzah Salim and Raj Yadav, 2012 [3]; Mohammad Fawzi Shubita and Jaafer Maroof Alsawalhah, 2012 [4]; Tran Thuy Minh Chau, 2018 [5]; Tran Thi Bich Ngoc, Nguyen Viet Duc, and Pham Hoang Cam Huong, 2017 [14]. For control variables, only SIZE variable has a positive effect on ROA, all the remaining variables (including GROWTH and TANG) have no effect on ROA.

The results found above indicate that the financial managers of these banks need to be aware of the role of capital structure in their profitability. It's possible to see that the debt of these banks accounts for a very high proportion of the total capital (over 90%) and the research results also show that the use of debt has an adverse impact on the bank's profitability. Therefore, the managers should consider reducing their debt ratio.

In addition, these banks also need to make better use of the firm's size to improve their profitability because the size of the business positively affects the profitability.

Results from this study also imply that to evaluate the profitability of a bank, investors need to pay attention to the capital structure in addition to other factors in the market. Therefore, they will have a more accurate investment analysis.

## REFERENCES

- [1] Brounen D., Real estate securitization and corporate strategy, Amsterdam : University of Amsterdam, 2003
- [2] Tailab, M. M. K., The effect of capital structure on profitability of energy American firms, *International Journal of Business and Management Invention*, Vol. 3, No. 12, pp. 54–61, 2014.
- [3] Mahfuzah Salim and Raj Yadav, Capital Structure and Firm Performance: Evidence from Malaysian Listed Companies, *Procedia – Social and Behavioral Sciences*, Vol. 65, No. 3, pp. 156–166, 2012.
- [4] Mohammad Fawzi Shubita and Jaafer Maroof Alsawalhah, The Relationship between Capital Structure and Profitability, *International Journal of Business and Social Science*, Vol. 3, No. 15, pp 105–112, 2012.
- [5] Tran Thuy Minh Chau, The effect of capital structure on the profitability of Vietnamese listed firm, Master's thesis, Danang University of Economics, 2018.
- [6] Abor, J., The effect of capital structure on profitability: An empirical analysis of listed firms in Ghana, *Journal of Risk Finance*, Vol. 6, No. 5, pp. 438–445, 2005.
- [7] Gill Amarjit, Nahum Biger, and Neil Mathur, The effect of capital structure on profitability: Evidence from the United States, *International Journal of Management*, Vol. 28, No. 4, Part 1, pp. 3–15, 2011.
- [8] Suleiman M. Abbadi and Nour Abu Rub, The Effect of Capital Structure on the Performance of Palestinian Financial Institutions, *British Journal of Economics, Finance and Management Sciences*, Vol. 3, No. 2, pp. 92–101, 2012.
- [9] Singh Narinder Pal and Bagga Mahima, The Effect of Capital Structure on Profitability: An Empirical Panel Data Study, *Jindal Journal of Business Research*, Vol 8, No. 1, pp. 65–77, 2019.
- [10] Modigliani, F., and Miller, M. H., Corporate income taxes and the cost of capital: A correction. *American Economic Review*, Vol. 53, No. 3, pp. 433–443, 1963.
- [11] Tran Thi Bich Ngoc, Nguyen Viet Duc, and Pham Hoang Cam Huong, The impact of capital structure on the performance of joint stock companies in Thua Thien Hue province, *Journal of Economics and Management Science*, Vol. 4, pp. 1–14, 2017.
- [12] Muhammad Ali Jibran Qamar , Umar Farooq , Hamayun Afzal, and Waheed Akhtar, Determinants of Debt Financing and Their Moderating Role to Leverage-Performance Relation: An Emerging Market Review, *International Journal of Economics and Finance*; Vol. 8, No. 5, 2016.
- [13] George Akerlof, Michael Spence and Joseph Stiglitz, The Market for "Lemons": Quality Uncertainty and the Market Mechanism, *The Quarterly Journal of Economics*, Volume 84, No. 3, pp. 488–500, 1970.
- [14] Pettit, R.R. & Singer, R.F., Small business finance: a research agenda, *Financial Management*, Vol. 14, No. 3, pp.47–60, 1985.
- [15] Myers, S.C. & Majluf, N.S., Corporate financing and investment decisions when firms have information that investors do not have, *Journal of Financial Economics*, Vol. 13, No. 2, pp. 187–221, 1984.
- [16] Modigliani, F. and Miller, M., The cost of capital, corporation finance and the theory of investment, *The American Economic Review*, Vol 48, No. 3, pp. 261–297, 1958.
- [17] Pham Thi Van Trinh, Panel data regression with DGMM method: Analytical technique in an empirical research, Vol. 54, No. 8, pp. 35–36, 2016.
- [18] Phan Thanh Hiep, Effect of capital structure on the performance of industrial manufacturing enterprises, Vol. 54, No. 6, pp. 9–13, 2016.

# Carboxymethyl Cellulose /Aloe Vera Gel Edible Films For Food Preservation

Hung Ngoc Nguyen

Department of Food Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
15116096@student.hcmute.edu.vn

Khoe Dang Dinh

Department of Food Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
15116100@student.hcmute.edu.vn

Linh T. K. Vu

Department of Food Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
linhvtk@hcmute.edu.vn

**Abstract**— Carboxymethyl cellulose (CMC) - based edible coatings have been widely investigated and used for fruits and vegetables preservation. In this study, the CMC edible films incorporated with *Aloe vera* gel (AVG) powder at different proportions (5/0, 5/1, 4/1, 3/1, and 3/2) were prepared and characterized to evaluate their applicability in food preservation. FT-IR spectra of the prepared films indicated that no major chemical differences in comparison to individual material. Mechanical strength test showed that as the *Aloe vera* gel proportion increased, the tensile strength and puncture resistance of the CMC/AVG films decreased, while the elongation at break significantly increased. The addition of *Aloe vera* gel not only lowered the water vapor permeability and but also increased the water solubility of CMC based films. Especially, all the tested CMC/AVG films were impermeable to oil. These results suggested that CMC/AVG edible films are applicable to postharvest preservation of some fruits such as banana, custard apple, and dragon fruit. In addition, CMC/AVG films are also suitable for packaging high fat-containing food products, such as the manufacturing of oil packets of instant noodles due to their high oil resistance and high dissolvability in hot water.

**Keywords**— carboxymethyl cellulose, aloe vera gel, edible films/coatings, food preservation

## I. INTRODUCTION

Plastic-based environmental pollution has led to the development and usage of biodegradable or edible packaging over the past few decades to reduce packaging waste [1, 2]. Biodegradable films and coatings can be produced from biodegradable materials such as proteins, polysaccharides, lipids, and other food-grade additives [3]. Among these biopolymers, carboxymethyl cellulose (CMC) – a derivative of cellulose – is the commonly used packaging material because of its high solubility in hot and cold water, ability to form nontoxic and flexible films, high water holding capacity, good biodegradability and low cost [4]. CMC has been applied in several edible film formulations for extending shelf-life of different types of fruits such as mandarin, orange and grapefruit [5], squash [6], and also lipid food [7]. However, some disadvantages of CMC coating are its high-water vapor permeability, low mechanical strength, and lack of antimicrobial property [8]. Hence, CMC coatings were incorporated with other components such as starch, *Dianthus barbatus* essential oil, or even *Aloe vera* gel to improve its mechanical and protective properties [8-10].

*Aloe vera* gel (AVG) is a colorless mucilaginous fraction obtained from the parenchymatous tissues of *Aloe vera* leaves [11]. It has been well reported that many of the medicinal

effects of AVG were attributed to its polysaccharide components such as pure mannan, acetylated mannan, acetylated glucomannan (acemannan), glucogalactomannan [11, 12]. In addition, due to its hygroscopic property, anti-fungal activity, biodegradability and biochemical properties, AVG has recently been used in the formulations of edible preservative coatings to prevent moisture loss, retard weight loss, control respiration exchange, delay oxidation and reduce microorganism proliferation of some types of fruits [13, 14]. With the benefits of both CMC and AVG, the combination of CMC and AVG can create a potential edible coating with good preservative and mechanical properties. Besides, the good solubility of CMC in cold water is also beneficial for the synthesis of CMC/AVG membranes because high heat treatment could damage the biological properties or active compounds in AVG. Santoso and Rahmat (2013) combined CMC with *Aloe vera* gel (AVG) to form an edible coating used in preserving fresh tomatoes [9]. Results showed that after 15 days, CMC-AVG coated tomatoes had higher freshness, such as maintaining their texture firmness and reducing weight loss, as compared to uncoated samples. However, to our best knowledge, the properties of CMC/AVG edible film have not been substantially studied. Therefore, the main goal of this study is to focus on the synthesis and characterization of CMC edible film incorporated with AVG to determine the quality of CMC/AVG film and evaluate its applicability in food preservation.

## II. MATERIALS AND METHODS

### A. Materials

*Aloe vera* gel powder 200x (*Aloe barbadensis*) was manufactured by Herbs & Crops Overseas (India) and was distributed by HNC INC (USA). Carboxymethyl cellulose (CMC) and glycerol were purchased from Shandong Yulong Cellulose Technology Co., Ltd. (China) and Xilong Scientific Co., Ltd. (China), respectively. Other chemicals and reagents such as sodium chloride, glycerol, silicagel were supplied by Hoa Nam company (Vietnam), and were of analytical grade.

### B. Preparation of CMC/AVG films

CMC/AVG films were prepared using the solvent-casting process [15]. Solutions used to prepare films (2.5% w solid/v) were prepared by dispersing CMC and AVG powder at different concentrations in distilled water (Table 1). Glycerol (20% w/w solid) was added as the plasticizer. The film solutions were mildly heated at 50 °C and stirred at 200 rpm for 30 min. The solutions were then sonicated for 15 min to remove air bubbles. The film-forming solutions (60 mL) were casted into Teflon dishes ( $\varnothing = 7$  inch). The films were left to

dry in drying oven (U1, Memmert, Germany) at 50 °C for 24 h. Prior to the test, the film samples were conditioned at  $30 \pm 2$  °C and  $75 \pm 5\%$  relative humidity (RH) for 24 h.

TABLE I. COMPOSITION OF CMC, AVG POWDER IN 100 mL FILM SOLUTION

Film	Composition (g/100 mL water)		
	CMC	AVG	Glycerol
CMC	2.5	0	0.5
CA51	2.084	0.416	0.5
CA41	2	0.5	0.5
CA31	1.875	0.625	0.5
CA32	1.5	1	0.5

### C. Film characterization

#### 1) Film thickness

The film thickness at moist state (conditioned at 75% RH for 24 h) was measured using a digital caliper (D0022-06; C-Mart, Taiwan) at nine different positions on the film. The results were expressed as a mean of the measurements  $\pm$  standard deviation (SD).

#### 2) Moisture absorption

Moisture absorption was determined according to the method of Angles and Dufresne [16]. The film samples (20 x 20 mm) were first dried at 60 °C until constant weight ( $W_i$ ). The dried film samples were then conditioned at  $30 \pm 2$  °C in a desiccator containing sodium chloride saturated solution to ensure  $75 \pm 5\%$  RH. After 24 h, the samples were weighed using a four-digit weighing balance ( $W_f$ ). The water uptake of the films was calculated by Eq. (1):

$$MA = \frac{W_f - W_i}{W_f} \times 100 \quad (1)$$

#### 3) Film light transmission

Film light transmission and transparency were determined by using a spectrophotometry (UH5300 Spectrophotometer, Hitachi, Japan). Rectangular film samples (10 mm x 35 mm) were placed into a glass cell and their light barrier properties were measured at wavelength between 200 and 800 nm, using air as a control.

#### 4) Chemical structure characterization

The chemical structure of CMC/AVL films were analyzed using a Fourier Transform Infrared (FTIR) system (FTIR-8400S, Shimadzu, Japan) at the Research and Development Center for Radiation Technology (VINAGAMMA). The samples (CMC powder, AVG powder and CMC/AVG film powder) were scanned in the range of 400 – 4000  $\text{cm}^{-1}$  with a resolution of 4  $\text{cm}^{-1}$ . The obtained spectrum was analyzed using the software SigmaPlot (Version 10.0).

#### 5) Mechanical properties

Mechanical properties were obtained using a Brookfield CT3TM Texture Analyzer (USA). Prior to test, the film samples were conditioned at  $30 \pm 2$  °C and  $75 \pm 5\%$  RH for 24 h.

##### a) Tensile strength

Tensile strength (TS, MPa) and elongation at break (E, %) of the films were determined according to ASTM standard method D882-10 [17]. The film strips (15 x 100 mm) were

tested at room temperature, using a gauge length of 50 mm and a crosshead speed of 1 mm/s. Tensile strength (TS, MPa) and elongation at break (E, %) were determined by Eq. (2) and Eq. (3):

$$TS = \frac{F_{\max}}{A} \quad (2)$$

$$E = \frac{L - L_0}{L_0} \times 100 \quad (3)$$

where F is the maximum force at break (N); A represents the cross-sectional area of the film strip ( $\text{m}^2$ ), L corresponds to the final length of the specimen at rupture (mm), and  $L_0$  represents the initial length of the specimen prior the test (mm) [15].

##### b) Puncture resistance

Puncture resistance was determined through the method of Preis et al. [18]. A minimum of five specimens from each sample film were prepared and cut into square pieces of 3 cm x 3 cm. Films were fixed by screws between two plates with a cylindrical hole of 10 mm diameter. Four pins stabilized the plates that were placed centrically under the punch of TA – MTP - 4R probe. The probe was adjusted to move forward with a velocity of 1.0 mm/s. Measurement started when probe had contact to the sample surface (triggering force = 10g). The probe moved on at constant speed until the film broke apart. The applied force and displacement (penetration depth) were recorded. Puncture strength (PS) was determined by the Eq. (4):

$$\text{Puncture strength} = F/A \quad (4)$$

where F describes the maximum applied force recorded during the strain (N); and A is the probe contact area ( $A = 12.56 \text{ mm}^2$ );

Elongation to break was determined by the Eq. (5):

$$\text{Elongation to break} = \left( \frac{\sqrt{a'^2 + b^2} + r}{a} - 1 \right) \times 100 \quad (5)$$

where r is the radius of the probe, a is the radius of the film in the sample holder opening,  $a'$  is the initial length of the film sample that is not punctured by the probe ( $a' = a - r$ ), b is the penetration depth/vertical displacement by the probe [18].

##### 6) Water vapor permeability (WVP)

WVP of CMC/AVG films was determined gravimetrically using a modified ASTM E96-95 [19]. The test film was sealed to a glass bottle containing 65 g silicagel to produce 0% RH below the film. The bottle was placed in a desiccator maintained at 75% RH with saturated sodium chloride. The water vapor transferred through the film and absorbed by the desiccant was determined by measuring the weight gain. WVP was calculated by Eq. (6):

$$WVP = \frac{w \cdot x}{t \cdot A \cdot P \cdot \Delta RH} \quad (6)$$

where WVP is the rate of water vapor transmission ( $\text{g} \cdot \text{s}^{-1} \cdot \text{m}^{-1} \cdot \text{Pa}^{-1}$ ); x is the film average thickness (m); P is the partial water vapor pressure at test temperature (3.169 kPa), A = permeation area (bottle mouth area,  $\text{m}^2$ ),  $\Delta RH$  is the difference in relative humidity (0.75).

##### 7) Solubility in water

The soluble matter in water of the films was measured according to Bierhalz et al. [20]. Square pieces of sample film (3 cm x 3 cm) were dried to constant mass at 60 °C for 24 h



and weighed ( $W_i$ ). The films were then immersed in 20 mL of distilled water at 25 °C. After 24 hours, the films were taken out to determine the final dry matter ( $W_f$ ) in the same drying condition (60 °C/ 24 h). The water solubility ( $S_w$ , %) was calculated by Eq. (7):

$$S_w = \frac{W_i - W_f}{W_i} \times 100 \quad (7)$$

#### 8) Film permeability to oil

Film permeability to oil was determined according to Wang et al. [21]. Whatman filter papers were dried to a constant weight ( $w_i$ ). A trimmed test film disc was placed on an individual filter paper and twenty-five drops of oil (0.6625 g) was evenly spread onto the film surface without exceeding the edge of the film. The test film was held for 24 h. The oil and test film were removed from the filter paper and the filter paper was reweighed ( $w_f$ ). Film permeability to oil was calculated by Eq. (8):

$$PO = \frac{w_f - w_i}{0.6625} \times 100 \quad (8)$$

#### 9) Statistical analysis

The data was presented as mean  $\pm$  standard deviation (SD). The statistical analysis of the data was performed using the one-way analysis of variance (ANOVA) and the comparisons between two means were done through the t-test. Statistical significance was considered at  $p < 0.05$ .

### III. RESULTS AND DISCUSSION

#### A. Film thickness

CMC/AVG thin films were prepared by the solvent – casting technique using different CMC/AVG ratios (Table 1). Glycerol was added to improve the flexibility and malleability of the films [15]. The CMC/AVG films presented a smooth surface and thicknesses were in the range of 93 – 103  $\mu\text{m}$ , significantly higher than that of the control film CMC (Table 2). Thinner thickness could be attributed to the formation of a more compact arrangement between the same polymer molecules within CMC film [23]. The AVG powder, on the other hand, contains a mixture of polysaccharides including glucomannan, galactane, arabinan and pectic compounds; these components have higher affinity for water, providing more free hydroxyl groups for the films than the CMC molecules alone [24]. The higher hydrophilic property of AVG hence increased the water absorption capacity of the CMC/AVG films, resulting in a significant increase in the film thickness at moist state.

TABLE II. PROPERTIES OF CMC/AVG FILMS

Film	Thickness ( $\mu\text{m}$ )	Moisture absorption (%)	Water solubility (%)
CMC	65.9 $\pm$ 5.4	17.3 $\pm$ 2.8	76.5 $\pm$ 3
CA51	89.6 $\pm$ 5.0	21 $\pm$ 2.5	94.3 $\pm$ 4.6
CA41	100 $\pm$ 5.9	22.7 $\pm$ 3.7	100
CA31	90 $\pm$ 4.1	33.4 $\pm$ 5.0	100
CA32	93.8 $\pm$ 6.4	35.6 $\pm$ 4.4	100

#### B. Moisture absorption

Moisture absorption is a parameter related to the total void volume occupied by water molecules in the microstructure network of the film during conditioning at 75% RH. As can be seen in Table 2, the increase in AVG content caused a

significant increase in moisture absorption. This was attributed to the hydrophilic property of the *Aloe vera* gel [15].

#### C. Film light transmission

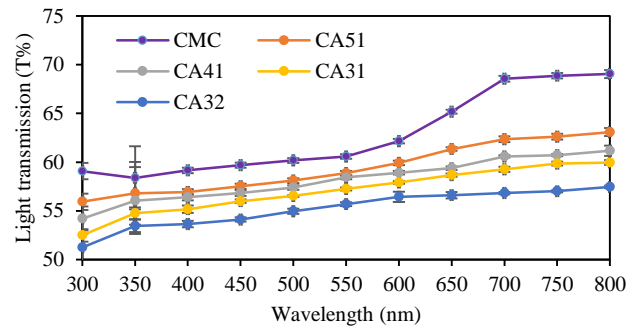


Fig. 1. Light transmission properties of the CMC/AVG films.

The light transmission and transparency are important properties of food packaging materials. Results in Figure 1 show that all the films exhibited weak light transmission in the ultraviolet light region (300 and 350 nm). As the wavelength increased in the visible range of 350 – 800 nm, the light transmission of CMC/AVG films increased. In the case of the control film (CMC), the film light transmission increased until a maximum of 69  $\pm$  0.9%. However, the incorporation of AVG within the CMC films caused a significant decrease in the light transmission at each wavelength. This reduction could be due to the high retention of the light. In particular, the light transmission of CMC/AVG films reached maximum values of 63  $\pm$  0.4% (CA51), 61  $\pm$  0.3% (CA41), 60  $\pm$  0.3% (CA31) and 57.4  $\pm$  0.2% (CA32). Lower light transmission of CMC/AVG films in the visible range could be beneficial in food packaging to retard deterioration.

#### D. Chemical structure characterization

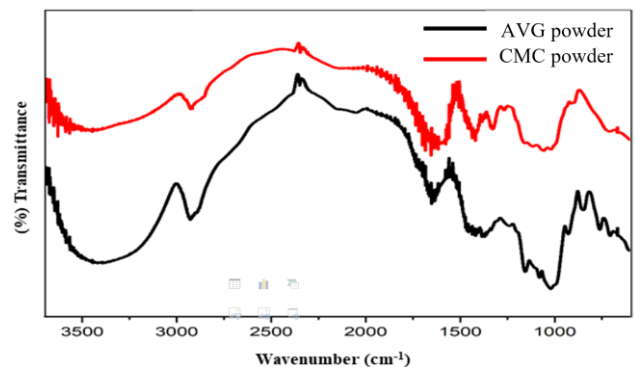


Fig. 2. Infrared spectra of CMC and AVG powder.

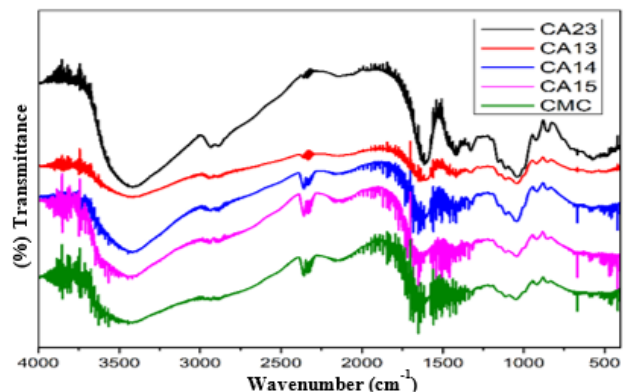


Fig. 3. Infrared spectra of CMC/AVG films.

FT-IR analysis of CMC/AVL films was performed to detect the chemical groups present in their chemical structure and to investigate the possible interactions between CMC and AVG. The infrared spectra of CMC powder and AVG powder are shown in Figure 2.

As can be seen in Figure 2, CMC powder presents characteristic absorption bands in FTIR such as hydroxyl groups (-OH stretching) at  $3454\text{ cm}^{-1}$ , hydrocarbon groups (-CH<sub>2</sub> scissoring) at  $1424\text{ cm}^{-1}$ , carbonyl groups (-C=O) at  $1599.8\text{ cm}^{-1}$ , ether groups (-O-) at  $1058\text{ cm}^{-1}$  and also C-H stretching vibration at  $2924\text{ cm}^{-1}$  [25, 26].

AVG also presents a large absorption band at  $3410\text{ cm}^{-1}$ , corresponding to the stretching vibration of -OH group [27]. The absorption peak at  $2928.7\text{ cm}^{-1}$  is attributed to the -CH vibration [15, 27]. In addition, AVG presents two characteristic absorption bands at around  $1641\text{ cm}^{-1}$  and  $1418\text{ cm}^{-1}$ , corresponding to the asymmetric and symmetric stretching vibration of COO groups, respectively [28]. Peak at  $1255\text{ cm}^{-1}$  refers to C-O-C stretching of acetyl group [27]. It was suggested that the presence of acetyl groups is necessary for biological activity of acetylated glucomannan, possibly because they cover a number of hydrophilic hydroxyl groups, making the molecule more able to cross hydrophobic barriers in the cell [24, 29].

The FT-IR spectra of the CMC/AVG films are shown in Figure 3. There is a broad envelope between  $3700$  and  $3300\text{ cm}^{-1}$ , corresponding to the -OH stretching. The -OH groups are hydrogen bonded [30]. Hydrogen bonds were formed during the formation of films. They are derived from -OH groups of CMC molecules and polysaccharides found in AVG. It is also possible to observe a weak absorption band at  $1600\text{ cm}^{-1}$ , probably related to the C-O and C=O linkages, from the aldehyde/ketone and carboxylic acid groups [15]. Furthermore, absorption bands at  $1300\text{ cm}^{-1}$  and  $1100\text{ cm}^{-1}$  were also detected, probably due to the presence of acetyl vibration and C-O-C groups [27]. FT-IR analysis shows that the preparation of CMC/AVG films retained bioactive compounds in the *Aloe vera* gel. In addition, the formation of hydrogen bonds confirmed that there were chemical interactions between CMC and AVG in the preparation of CMC-based films.

## E. Mechanical properties

### 1) Tensile strength

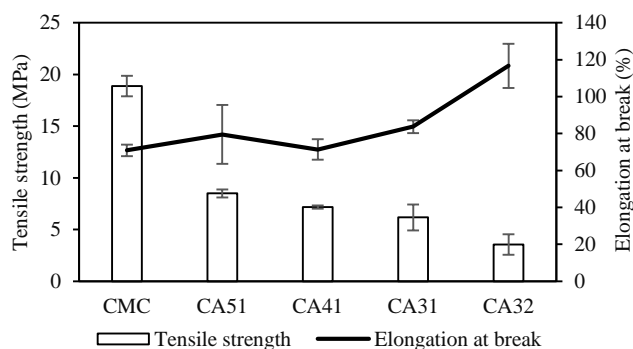


Fig. 4. Tensile Strength and Elongation to Break of CMC/AVG films

Food packaging should exhibit good mechanical properties that maintain their integrity and flexibility during handling and storage [15]. The mechanical properties of the

CMC/AVG films were evaluated regarding their tensile strength and elongation at break. Tensile strength is the largest stress required to break the film, while elongation is the degree to which film specimen can stretch before breaking [31]. As can be seen in Figure 6, the addition of AVG to the films significantly reduced the tensile strength from  $18.48\text{ MPa}$  (film CMC) to  $3.56\text{ MPa}$  (film CA32). On the other hand, the elongation at break of the CMC/AVG films was quite high, ranging from  $70.9\%$  (CA41) to  $116.6\%$  (CA32) (Figure 4). There was no significant difference in the elongation at break of the films CMC, CA51, and CA41. However, films CA31 and CA32 expressed higher elongation at break as compared to that of film CMC ( $83.7\%$  and  $116.6\%$  versus  $70.9\%$ ). This showed that AVG played a role as a plasticizer in the CMC films, resulting in a change in the extensibility of the film. AVG might have reduced the intermolecular force and increased the mobility of the polymeric chain. This is because the hydrophilic constituents in the AVG fit easily into CMC network, and by forming hydrogen bonds, CMC-CMC interaction could be reduced, resulting in better plasticity of the films [2]. The obtained result was in agreement with those reported by Pereira et al. [15] and Chin et al. [2] in which the *Aloe vera* gel improved the flexibility of alginate and gelatin films, respectively. The elongation at break of the CMC/AVG films in this study was also higher than that of alginate/AVG films ( $30.2\text{--}46.7\%$ , [15]) and gelatin/aloe gel films ( $1.47\text{--}1.95\%$ , [2]). Higher elongation at break indicates the films are more flexible when subjected to tension and mechanical stress [33].

### 2) Puncture strength

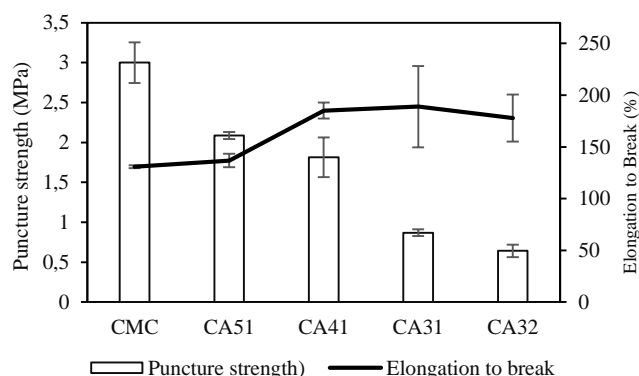


Fig. 5. Puncture strength and Elongation to break of CMC/AVG films.

Puncture strength is a measure of the maximum force or energy required to penetrate a material. The puncture force and elongation to break of CMC/AVG films are shown in Figure 5. It can be easily noticed that as the concentration of AVG in the composite films increased, the puncture strength of CMC/AVG films significantly decreased (from  $3.0\text{ MPa}$  to  $0.64\text{ MPa}$ ). The elongation to break of the CMC/AVG films also increased when increasing AVG concentration, from  $136.8\%$  (film CMC) to  $177.9\%$  (film CA31). High elongation to break of films CA32 and CA31 may be related to the high water absorption capacity of AVG components as discussed above.

To conclude, the addition of AVG to the CMC film reduced the tensile strength and penetration resistance while increasing the elongation at break as well as the flexibility of the CMC/AVG membranes.

#### F. Water vapor permeability (WVP)

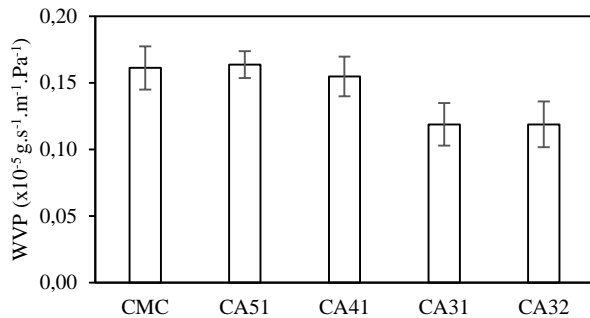


Fig. 6. Water Vapor Permeability (WVP) of CMC/AVG films.

WVP is an important criterion of the food packaging material, defined as the transmission rate of water to the driving force of vapor pressure [34]. Interactions between the wrapped product and the environment should be minimized to extend the shelf life of the products [2]. Results in Figure 6 show that, as the *Aloe vera* gel concentration in CMC films increased, the WVP of the CMC/AVG films decreased from  $0.161 \times 10^{-5} \text{ g.s}^{-1}.\text{m}^{-1}.\text{Pa}^{-1}$  (film CMC) to  $0.119 \times 10^{-5} \text{ g.s}^{-1}.\text{m}^{-1}.\text{Pa}^{-1}$  (films CA31 and CA32). This reduction could be explained by the formation of cross linkages between CMC and polysaccharides in the AVG, which reduced the free space available for the movement of water molecules [2]. In addition, due to hydrophilic property AVG, water molecules can be easily absorbed into the AVG-containing films and trapped there, resulting in a decrease in WVP with increasing AVG concentration in the films. Similar result was observed in WVP of chitosan/*Aloe vera* films [35], gelatin/*Aloe vera* films [2].

#### G. Water solubility

Solubility of a food packaging film is an important characteristic especially for food products with high moisture content [2]. Results in Table 2 show that films CA32, CA31 and CA41 were completely soluble in water. When the *Aloe vera* gel content in the films decreased, the solubility of the film significantly decreased. An explanation for this observation would be the high solubility and hydrophilicity of AVG [2], film samples containing higher AVG concentration would be able to absorb water faster. Therefore, CMC molecules in the films would be hydrated more quickly, contributing to the increasing in solubility.

In general, the solubility of CMC/AVG films was still high because the two main components CMC and AVG are both hydrophilic [2, 10]. Therefore, CMC/AVG films would not be suitable for food products with high moisture and wet surface such as meat product, fresh-cut fruit, and seafood. With water-soluble property, the CMC/AVG films could be used in packaging of food products with low water activity such as dried vegetables and fruits, or in coating of fruits such as banana, apple, custard apple.

#### H. Film permeability to oil

According to Figure 7, the oil permeability of the CMC/AVG films decreased as the *Aloe vera* concentration in CMC films increased. However, there was no significant difference in oil permeability between films CMC, CA51 and CA41 and between films CA32 and CA31. Besides, it can be seen that the CMC/AVG films had very low oil permeability (from 0.01% to 0.07%), or it can be inferred that they are

impermeable to oil. This can be explained by the hydrophilic nature of both of CMC and AVG, making CMC/AVG films resistant to non-polar oil molecules. Similar results were obtained in the study by Wang et al. (2007) in which the edible films produced from proteins or polysaccharides were impermeable to oil [21]. This attribution of CMC/AVG films makes them suitable for packaging of fatty food products.

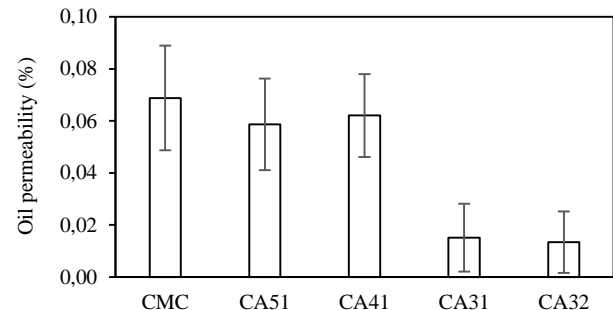


Fig. 7. Oil permeability of CMC/AVG films

#### IV. CONCLUSION

The findings suggested that the incorporation of *Aloe vera* gel into CMC films increased the film thickness, water solubility and decreased the film light transmission and water vapor permeability. The technological properties of the CMC/AVG films were greatly influenced by the strong hydrophilicity of the polysaccharides contained in AVG. The composite films had lower tensile strength and puncture strength compared to the control film. However, plasticity of the CMC/AVG films were significantly improved. Especially, the addition of AVG into the CMC films improved film impermeability to oil, making them more suitable for food packaging of fatty food products. For instance, CMC/AVG film can be used as the oil package in cup noodles due to its high oil resistance and high dissolvability in hot water.

#### REFERENCES

- [1] A. Jiménez, M. J. Fabra, P. Talens, and A. Chiralt, "Edible and Biodegradable Starch Films: A Review," Food and Bioprocess Technology, vol. 5, no. 6, pp. 2058-2076, 2012.
- [2] S. Sui Chin, F. Han Lyn, and Z. A. Nur Hanani, "Effect of Aloe vera (*Aloe barbadensis* Miller) gel on the physical and functional properties of fish gelatin films as active packaging," Food packaging and shelf life, Article vol. 2017 v.12, pp. pp. 128-134, 2017-06, 2017.
- [3] A. U. Malik, F. Siddiq, and M. Siddiq, "Packaging of Fresh Mangoes and Processed Mango Products," in Handbook of Mango Fruit: Production, Postharvest Science, Processing Technology and Nutrition, M. Siddiq, Jeffrey K. Brecht, and J. S. Sidhu, Eds.: John Wiley & Sons Ltd, 2017, pp. 131-149.
- [4] M. Tabari, "Investigation of Carboxymethyl Cellulose (CMC) on Mechanical Properties of Cold Water Fish Gelatin Biodegradable Edible Films," Foods, vol. 6, no. 6, May 27 2017.
- [5] H. Arnon, Y. Zaitsev, R. Porat, and E. Poverenov, "Effects of carboxymethyl cellulose and chitosan bilayer edible coating on postharvest quality of citrus fruit," Postharvest Biology and Technology, vol. 87, pp. 21-26, 2014.
- [6] A. G. Ponce, S. I. Roura, C. E. del Valle, and M. R. Moreira, "Antimicrobial and antioxidant activities of edible coatings enriched with natural plant extracts: In vitro and in vivo studies," Postharvest Biology and Technology, vol. 49, no. 2, pp. 294-300, 2008/08/01/ 2008.
- [7] S. Ganiari, E. Choulitoudi, and V. Oreopoulou, "Edible and active films and coatings as carriers of natural antioxidants for lipid food,"

- Trends in Food Science & Technology, vol. 68, pp. 70-82, 2017/10/01/ 2017.
- [8] M. Mohammadi, M. H. Azizi, and A. Zoghi, "Antimicrobial activity of carboxymethyl cellulose-gelatin film containing *Dianthus barbatus* essential oil against aflatoxin-producing molds," (in eng), Food science & nutrition, vol. 8, no. 2, pp. 1244-1253, 2020.
- [9] F. Santoso and V. A. Rahmat, "Safety and quality assurance of tomato using aloe vera edible coating," Acta Horticulturae, vol. 1011, pp. 133-140, 11/05 2013.
- [10] B. Ghanbarzadeh, H. Almasi, and A. A. Entezami, "Physical properties of edible modified starch/carboxymethyl cellulose films," Innovative Food Science & Emerging Technologies, vol. 11, no. 4, pp. 697-702, 2010/10/01/ 2010.
- [11] J. H. Hamman, "Composition and applications of Aloe vera leaf gel," (in eng), Molecules, vol. 13, no. 8, pp. 1599-616, Aug 8 2008.
- [12] C. Liu, Y. Cui, F. Pi, Y. Cheng, Y. Guo, and H. Qian, "Extraction, Purification, Structural Characteristics, Biological Activities and Pharmacological Applications of Acemannan, a Polysaccharide from Aloe vera: A Review," Molecules, vol. 24, no. 8, Apr 19 2019.
- [13] M. R. Sharmin, M. N. Islam, and M. A. Alim, "Shelf-life enhancement of papaya with aloe vera gel coating at ambient temperature," Journal of the Bangladesh Agricultural University, vol. 13, p. 131, 07/14 2016.
- [14] J. M. Valverde, D. Valero, D. Martínez-Romero, F. Guillén, S. Castillo, and M. Serrano, "Novel Edible Coating Based on Aloe vera Gel To Maintain Table Grape Quality and Safety," Journal of Agricultural and Food Chemistry, vol. 53, no. 20, pp. 7807-7813, 2005/10/01 2005.
- [15] R. Pereira, A. Carvalho, D. C. Vaz, M. H. Gil, A. Mendes, and P. Bartolo, "Development of novel alginate based hydrogel films for wound healing applications," Int J Biol Macromol, vol. 52, pp. 221-30, Jan 2013.
- [16] M. N. Anglès and A. Dufresne, "Plasticized Starch/Tunicin Whiskers Nanocomposite Materials. 2. Mechanical Behavior," Macromolecules, vol. 34, no. 9, pp. 2921-2931, 2001/04/01 2001.
- [17] A. D882-10, "Standard Test Method for Tensile Properties of Thin Plastic Sheeting," ASTM International, West Conshohocken, PA, www.astm.org., 2010.
- [18] M. Preis, K. Knop, and J. Breitzkreutz, "Mechanical strength test for orodispersible and buccal films," International Journal of Pharmaceutics vol. 461, no. 1-2, pp. 22-29, Jan 30 2014.
- [19] A. E. 95, "Standard Test Methods for Water Vapor Transmission of Materials," ASTM International, West Conshohocken, PA, www.astm.org., 1995.
- [20] A. C. K. Bierhalz, M. A. da Silva, and T. G. Kieckbusch, "Natamycin release from alginate/pectin films for food packaging applications," Journal of Food Engineering, vol. 110, no. 1, pp. 18-25, 2012.
- [21] L. Wang, L. Liu, J. Holmes, J. Kerry, and J. Kerry, "Assessment of film - forming potential and properties of protein and polysaccharide - based biopolymer films," International Journal of Food Science & Technology, vol. 42, pp. 1128-1138, 09/01 2007.
- [22] C. P. Chen, B.-e. Wang, and Y.-M. Weng, "Physiochemical and antimicrobial properties of edible aloe/gelatin composite films," International Journal of Food Science & Technology, vol. 45, pp. 1050-1055, 04/12 2010.
- [23] X. L. Chang, B. Y. Chen, and Y. M. Feng, "Water-soluble polysaccharides isolated from skin juice, gel juice and flower of Aloe vera Miller," Journal of the Taiwan Institute of Chemical Engineers, vol. 42, no. 2, pp. 197-203, 2011/03/01/ 2011.
- [24] S. Alizadeh Asl, M. Mousavi, and M. Labbafi, "Synthesis and Characterization of Carboxymethyl Cellulose from Sugarcane Bagasse," Journal of Food Processing & Technology, vol. 08, no. 08, 2017.
- [25] R. L. Wehling, "Infrared Spectroscopy," in Food Analysis., S. S. Nielsen, Ed.: Springer, Boston, MA, 2010.
- [26] N. R. Swami Hulle, K. Patruni, and P. S. Rao, "Rheological Properties of Aloe Vera (*Aloe barbadensis* Miller) Juice Concentrates," Journal of Food Process Engineering, vol. 37, no. 4, pp. 375-386, 2014.
- [27] R. Pereira, A. Tojeira, D. C. Vaz, A. Mendes, and P. Bártolo, "Preparation and Characterization of Films Based on Alginate and Aloe Vera," International Journal of Polymer Analysis and Characterization, vol. 16, no. 7, pp. 449-464, 2011/10/01 2011.
- [28] T. Reynolds and A. C. Dweck, "Aloe vera leaf gel: a review update," Journal of Ethnopharmacology, vol. 68, no. 1, pp. 3-37, 1999/12/15/ 1999.
- [29] T. A. Kuriakose et al., "Synthesis of stoichiometric nano crystalline hydroxyapatite by ethanol-based sol-gel technique at low temperature," Journal of Crystal Growth, vol. 263, no. 1, pp. 517-523, 2004/03/01/ 2004.
- [30] T. Janjarasskul and J. M. Krochta, "Edible packaging materials," Annu Rev Food Sci Technol, vol. 1, pp. 415-48, 2010.
- [31] M. H. Norziah and J. L. Wong, "Effects of Sucrose Palmitate on Physical and Mechanical Properties of Sago Starch-Gelatin Edible Films," in Gums and Stabilisers for the Food Industry 16: The Royal Society of Chemistry, 2012, pp. 257-268.
- [32] X. Li, A. Liu, R. Ye, Y. Wang, and W. Wang, "Fabrication of gelatin-laponite composite films: Effect of the concentration of laponite on physical properties and the freshness of meat during storage," Food Hydrocolloids, vol. 44, pp. 390-398, 2015.
- [33] S. Khoshgozaran-Abras, M. H. Azizi, Z. Hamidy, and N. Bagheripoor-Fallah, "Mechanical, physicochemical and color properties of chitosan based-films as a function of Aloe vera gel incorporation," Carbohydrate Polymers, vol. 87, no. 3, pp. 2058-2062, 2012.

# The Load-Bearing of Concrete Beams as the Steel Reinforcements Connected by the Coupler at a Cross-Section of a Beam

Thanh-Hung Nguyen  
Dept. of civil engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh, Vietnam  
nthung@hcmute.edu.vn

Phuong-Doanh Huynh  
Dept. of civil engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh, Vietnam  
hpdoanh.cema@gmail.com

Anh-Thang Le  
Dept. of civil engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh, Vietnam  
thangla@hcmute.edu.vn

**Abstract**—In this paper, the authors investigated the effect of the position and the splicing ratio of reinforcement bars by coupler on the behavior of reinforced concrete beams. The experimental samples gave the assessments with the reinforcement concrete beams with steel bars connected by coupler at the same cross-section. For supporting the results from experiments, a nonlinear finite element analysis was performed with the help of ABAQUS software. Comparison results of beams with/without coupler showed that the load-capacity and behavior of reinforced concrete beams with the couplers have relatively small changes compared to concrete beams without splice on steel bars.

**Keywords**—Reinforcement concrete beams, Reinforcement splice, Coupler, Reinforcement concrete beam model.

## I. INTRODUCTION

In the world, the steel reinforcement in reinforced concrete structure has been significantly improved. There are several methods of reinforcing steel to meet the requirements. Some studies in the world such as B. MacKay, D. Schmidt, and T. Rezansoff, (1998) [1] studied the influence of the overlapping ties in beams when subjected to loads. Rasha T.S. Mabrouk, Ahmed Mounir [2] has studied lap splices in reinforced concrete beams.

In order to limit the concrete locally crushed, the number of standards is limited to a reinforcement splices percentage that can be connected at a cross-section as NZS 3101-Part 1 (New Zealand Standard 2006) [3], limiting splices near the zone having the enormous tensile pressures such as ACI 318M-11 (ACI 2011) [4]. Ahmed El-Azab, Hatem M. Mohamed (2014) [5] investigated the influence of tension lap splice on the behavior of high strength concrete (HSC) beams. The study on beams with the connecting reinforcement is continued to date (Hardisty et al. 2015).

Therefore, assessing the effect of position and proportion of steel splices using couplers to the reinforced concrete beams is essential. The study will concentrate on the reinforcement concrete beam with the coupler on the same cross-section to have a more comfortable solution for splice positions of steel rebar on the construction.

## II. EXPERIMENTAL PROGRAM

### A. Test specimens

Experiment consisting of seven reinforced concrete beams prepared and experimented at the University of Technology and Education Ho Chi Minh City, Vietnam. The

beams were all in the size of 200x300x3300mm, with 4Ø16mm longitudinal steel, and steel stirrups of Ø6a150mm (Fig. 1). For beams named D1, D2, D3 have 100% longitudinal steel reinforcement bars splice connected with coupler. The positions of the splices were at 1/2, 1/3, 1/4 beam length. Beams named as D4, D5, D6 have the 75% of reinforcement bars splice connected with coupler, the splice positions at 1/2, 1/3, 1/4 beam length. The beam named as D7 was a reference beam without splices on reinforcement bars.

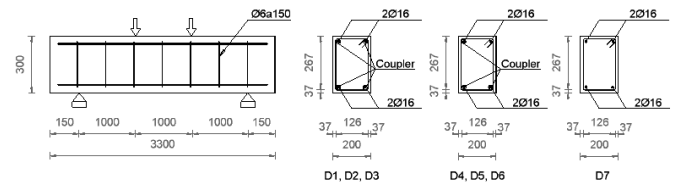


Fig. 1. Model of reinforced concrete beams.

### B. Materials used

The experiment used B20 concrete with a calculate strength of 11.5 MPa. The average cubical (15x15x15cm) compressive strength of concrete at the time of testing was 25.7 MPa. Table I shows the weights required to cast one cubic meter of concrete.

TABLE I. DESIGN OF THE CONCRETE MIX (PER M<sup>3</sup>)

Material	Weight
Coarse aggregate (Gravel) – m <sup>3</sup>	0.816
Fine aggregate (Sand) – m <sup>3</sup>	0.477
Cement (PCB40) - Kg	378
Water - Kg	182

Steel reinforcement quality was following TCVN 1651-2:2008, which has the specifications as in Table II.

TABLE II. SPECIFICATIONS OF BEARING STEEL

Grade of reinforcement	Yield strength (N/mm <sup>2</sup> )	Tensile strength (N/mm <sup>2</sup> )	Relative Elongation (%)
CB400-V	400	570	14

The using coupler, according to TCVN 8163: 2009 [6], having the characters shown in Table III. The material of the coupler had the steel strength higher than that of the steel strength of the primary reinforcement. Couplers were layout as Fig. 2.



TABLE III. SPECIFICATIONS OF A COUPLER.

Size (mm)	Outside diameter (mm) $\pm 1$	Length (mm) $\pm 1$	Pitch of thread (mm)	number of threaded rings (r) $\pm 0.5$	Yield stress (MPa)	Tensile strength (MPa)
16	25	42	2.5	16.8	$\geq 500$	$\geq 682.5$



Fig. 2. Steel spliced.

### C. Materials used

Four points bending test was applied to all the beams to ensure a pure bending in the mid zone. The testing machine consisted of a static hydraulic loading frame with a load cell used to apply the concentrated vertical load at the increment speed of 3.5 kN/s. The load from the testing machine was transferred through a stiffness steel beam placing onto the specimens to divide the load into two equally concentrated loads. A digital load indicator with a 1 kN accuracy was used to measure the applied load. At each load stage, beam deflection reading from LVDT was taken at the mid-span and the one-third of the beam. The using LVDT instrument had an accuracy of 0.01 mm. The strain of steel reinforcement was measured by strain gauges attached near the splice position, as shown in Fig.3.



Fig. 3. Strain gauges attached to the steel reinforcement

## III. EXPERIMENTAL RESULTS AND EVALUATION

After the process of the increasing load to bend the beams until the beams are destroyed, the empirical results show the relationship between the load and the deflection of the beams, the load and steel strain in the beams, the load capacity when cracks initially occur, and when beams destroyed, the crack pattern after beams destroyed. The assessment of the impact of splice varying with position and proportion to the behavior of reinforced concrete beams were explored.

### A. Deflection and load

The relationship between the bending load and the deflection at the steel coupler positions of the beam was demonstrated in the same chart with the reference beam for comparison, evaluation.

The measured positions of the beam deflection during the loading process were showed in Fig. 4, while Fig. 5 and Fig.

6 represent the load-deflection relationship of seven beams at position No. 1.

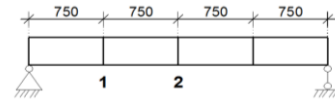


Fig. 4. Deflection measurement position

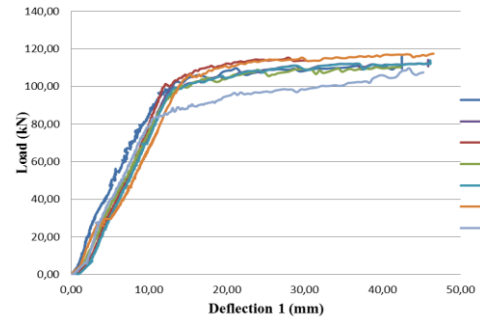


Fig. 5. Load deflection curves for beams in position No. 1

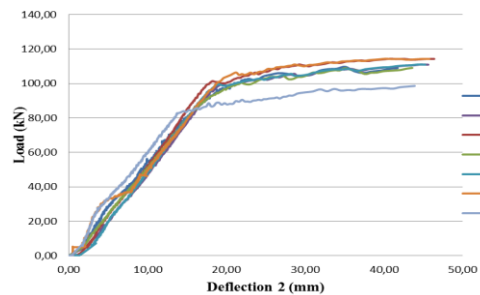


Fig. 6. Load-deflection curves for beams in position No. 2

Fig. 6 indicates that the loading capacity of concrete beams with the connected couplers up to 75% and 100% reinforcement area in a cross-section had a small difference at about 2% to 6%, respectively.

The couplers connect the reinforced steel, at positions 1/2, 1/3, 1/4 of beam length, reducing the loading capacity of beams in the phase with cracks and without cracks are relatively small. The average reduction was 11.5% and 12.5%, respectively. With an 11.5% reduction in the load-bearing capacity of the beams in the elastic period without cracks is not dangerous and can be acceptable.

### B. Strain and load

The relation between the load and steel strain of the beams was manifested on the same chart for comparison.

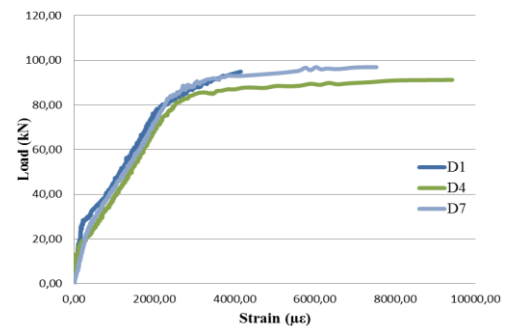


Fig. 7. The relationship between load and steel reinforcement strain of the D1, D4, and D7 beam

From Fig. 7, we can see, when the load under 80kN, reinforced in the working beam in the elastic domain, then the linear deform increases corresponding to each load level and reaches a value of about  $2250 \mu\epsilon$  at the 80kN load level. When the load exceeds 80kN, the reinforcement might work in the elastoplastic behavior, and the deformation of the steel reinforcement increases rapidly at the load level. In the elastic domain, the deformation under the loading of the beams having the steel couplers and the reference beam has a relatively small difference. This difference value is about 12%.



Fig. 8. The coupler of D1 and D4 beam after beams destroyed

After the beam was destroyed, the steel reinforcements were not moved from the coupler, and the coupler has no deformation phenomenon. Reinforcement and coupler were tightly connected almost absolutely as a consistent steel bar.

### C. Cracking load and failure load

Fig. 9 shows the initial cracking load and failure load of beams. In the elastic period, the initial cracking load of the test beams and the reference beams had a relatively small difference, about 4%. Therefore, almost coupler connecting the steel reinforcements did not affect the load-bearing capacity of the reinforced concrete beams. The failure load of the reference beam was higher than the beams having coupler; it is possible to see the behavior the reinforcement weakened by the coupler, reducing the load tolerance of beams.

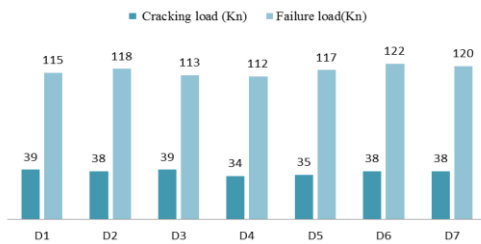


Fig. 9. Cracking load and failure load of beam samples

## IV. NUMERICAL SIMULATION

A three-dimensional nonlinear finite element analysis was undertaken to support the experimental results, with the help of the commercial finite element software, ABAQUS version 6.13[7]. ABAQUS is a complex finite element analysis program introduced with many material characteristics and parameters to reproduce high accuracy in calculations and provide comprehensive outputs concerning stress analysis.

The 3300x200x300 (mm) beams were modeled with two ends supported at a distance of 150mm from beam ends. A vertical load was applied in the middle of the beam. Tensile reinforcement was 2D16, while the compressive steel-

reinforced was 2D16. Rebar stirrups were D6@150 (AIII steel grade according to Vietnamese specification).

The C3D8R element type in the material library of the Abaqus software is used for the model. The C3D8R element is a 3D element with the eight linear nodes assigned to the concrete elements.

The reinforcement bars were modeled by the type of the truss element, T3D2. The main input parameter of this element type is the area of cross-section. The geometry of the cross-section does not need a specific definition. The coupler was modeled by the truss element, which is 4.2 cm in length, using the same material and has a 12% reduction area comparing to the longitudinal steel reinforcement.

Reinforcement and coupler are declared embedded into concrete with adhesion between reinforcement steel and concrete that is fixed.

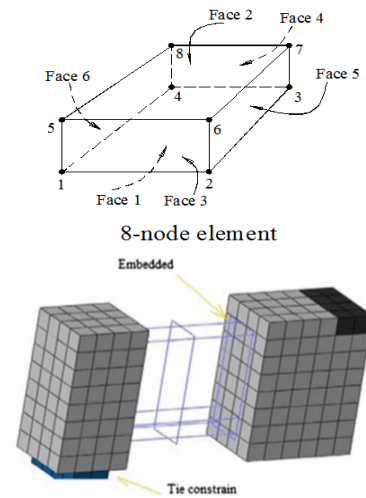


Fig. 10. Concrete model (C3D8R) and Embedded technique [2]

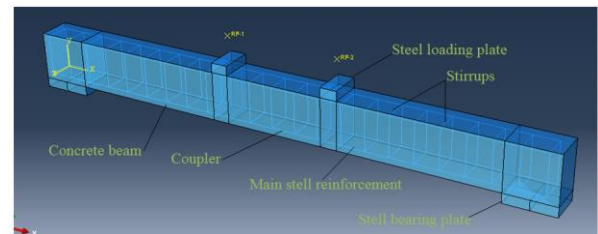


Fig. 11. Structural beams simulation

### A. Model materials in ABAQUS

In the study about the computational models of concrete beams done by Le Anh Thang, Nguyen Tat Thanh, Hoang Trong Quang [8], they concluded that the elastic – plasticizing model (used for reinforcement) combined with the model of Hsu-Hsu 1994 (used for concrete) for the simulation results in close to most experimental results. According to this result, the article uses two models of this material for simulation.

#### 1) Models of steel reinforcement material.

The selected model of reinforcement steel material defined in the ABAQUS environment was a simple elastoplastic law (IEPL). The Young modulus defined the line of the stress-strain relationship before the tensile yield strength of  $f_y$ . The critical tensile strength of steel was  $f_u$ . The model has been successfully applied in several studies of authors such as Ngo and Scordelis, Vebo, and Ghali.

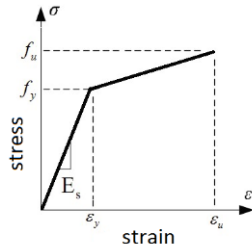


Fig. 12. Elastoplastic constitutive law of reinforcement

### 2) Models of concrete material.

The numerical model of concrete material proposed by Hsu–Hsu (1994). The complete stress-strain curve for concrete under compression was derived using the experimentally verified numerical method by Hsu and Hsu (1994). This model can be used to develop the stress-strain relationship under uniaxial compression up to  $0.3\sigma_{cu}$  of stress in the descending portion only using the maximum compressive strength ( $\sigma_{cu}$ ). Fig. 13 defines the ultimate compressive stress ( $\sigma_{cu}$ ), strain at  $\sigma_{cu}$  ( $\epsilon_0$ ), and the strain corresponding to the stress at  $0.3\sigma_{cu}$  in the descending portion ( $\epsilon_d$ ). A linear stress-strain relationship which obeys Hooke's law is assumed up to 50% of the ultimate compressive strength ( $\sigma_{cu}$ ) in the ascending portion. The numerical model by Hsu and Hsu (1994) is used only to calculate the compressive stress values ( $\sigma_c$ ) between the yield point (at  $0.5\sigma_{cu}$ ) and the  $0.3\sigma_{cu}$  in the descending portion using (1).

The numerical model by Hsu and Hsu (1994) is used only to calculate the compressive stress values ( $\sigma_c$ ) in the descending curve portion using (1).

$$\sigma_c = \left( \frac{\beta \cdot \left( \epsilon_c / \epsilon_0 \right)}{\beta - 1 + \left( \epsilon_c / \epsilon_0 \right)} \right)^\beta \cdot \sigma_{cu} \quad (1)$$

Where  $\epsilon_c$  is compressive stress values which depend on  $\sigma_c$  in curve stress-strain relationship, the parameter  $\beta$  and the strain at peak stress  $\epsilon_0$  is given (2), (3).

$$\beta = \frac{1}{1 - \left[ \sigma_{cu} / \epsilon_0 \cdot E_0 \right]} \quad (2)$$

$$\epsilon_0 = 8.9 \times 10^{-5} \cdot \sigma_{cu} + 2.114 \times 10^{-3} \quad (3)$$

The modulus  $E_0$  is given by (4). The  $\epsilon_d$  is the strain at  $\sigma_c = 0.8\sigma_{cu}$ .

$$E_0 = 1.2431 \times 10^2 \cdot \sigma_{cu} + 3.28312 \times 10^3 \quad (4)$$

The complete stress-strain curve for concrete under tension is derived using the experimentally verified numerical method by Hsu and Hsu (1994). Fig.13. Where  $\sigma_{t0}$  is the maximum tensile stress corresponding to critical tensile strain  $\sigma_{cr}$ .

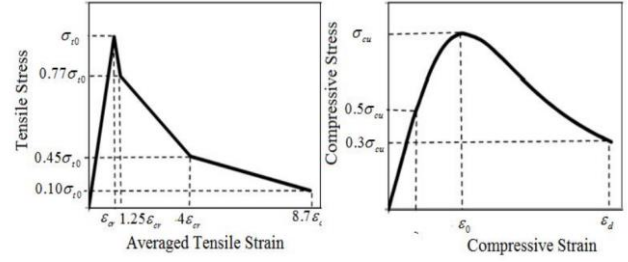


Fig. 13. Stress-strain curve for concrete under tension and compression by Hsu – Hsu

### 3) Calculation parameters for the model

The parameters for nonlinear analysis:

The material required parameters for nonlinear analysis are presented in Table.IV. The elements named as C3D8R in the material library of ABAQUS is used for concrete material models. The concrete beam is modeled in the form of a solid-element. The choice of the forming wire in ABAQUS is applied for the reinforced bars. The reinforcing bar is embedded in concrete.

TABLE IV. SPECIFICATIONS OF A COUPLER.

Concrete	$E_c$ (MPa)	$\nu_c$	$f_c$ (MPa)	$f_t$ (MPa)	
	26.8	0.2	30.31	3.06	
Steel	$E_s$ (MPa)	$\nu_c$	$f_y$ (MPa)	$f_u$ (MPa)	
	210	0.3	365	420	
Other parameters	$K_c$	$E$	$\sigma_{p0} / \sigma_{x0}$	$\Psi$	$\mu$
	0,667	0,1	1,16	300	0,00005

### Size mesh.

The mesh size affects convergence and the results of the analysis. Suitable element sizes are selected so that they do not affect analysis results. In the study, the selected mesh had an element size of 80mm (Mesh-80).

### B. Simulation results and comparison with experiments

The relationship between load and deflection in the middle of the beams is comparable to the experiment beam.

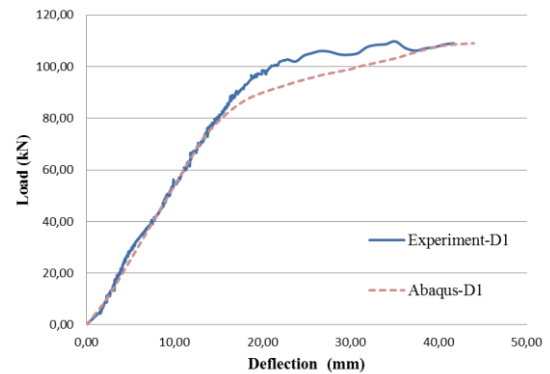


Fig. 14. Load-deflection curve of the D1 beam and model of D1

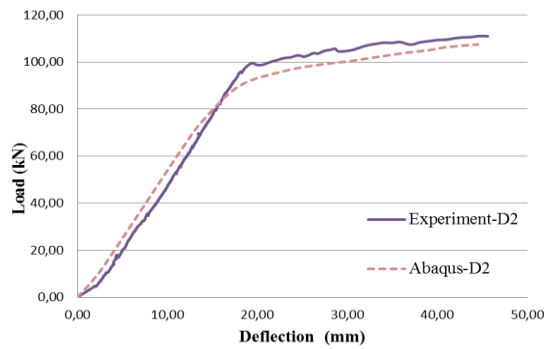


Fig. 15. Load-deflection curve of the D2 beam and model of D2

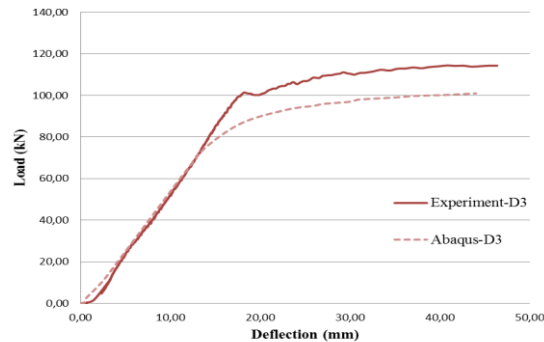


Fig. 16. Load-deflection curve of the D3 beam and model of D3

The relationship between load and strain in steel reinforcement of test beams is compared to the simulation beam in Fig. 17. It could be seen that they are nearly identical in shape.

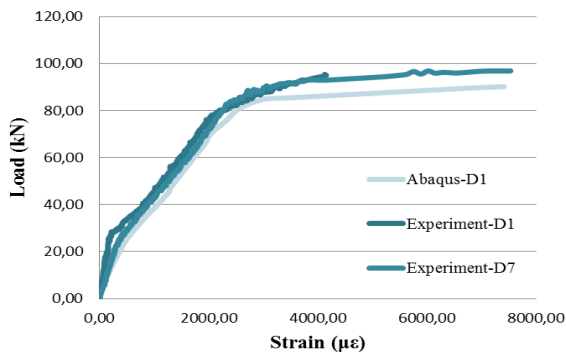


Fig. 17. Load – steel reinforcement strain curves of D1, D7 beam, and model of D1

In terms of the loading capacity, the curve lines of test and simulated beams had an average deviation of 2% as the beams working in the elastic behavior. As the beams working in the nonelastic behavior, the curves of simulated beams were less than those of test beams about 7%.

In terms of strain in steel, the curve lines of test and simulated beams had an average deviation of 5% as the beams working in the elastic behavior. For the nonlinear behavior, it was 10%. These tolerances are acceptable.

## V. CONCLUSIONS AND RECOMMENDATIONS

The study noticed the advantages of using coupler in the joint of reinforced concrete beams. Based on the

experiment results, the authors have some conclusions and recommendations following:

- Reinforced concrete beams having several reinforcing at a cross-section connected by the coupler affect the bearing capacity of the beam. However, the effect of couplers to beam loading-deflection is relatively low at about 11.5% for the elastic phase, 12.5% for the elastic stage with cracks, and this value is acceptable.
- When reinforcement steel is connected by coupler with the quantity of connects from 75% to 100%, at the positions of 1/2, 1/3, 1/4 of beam length, there is no change in the behavior of beams when subjected to static loads compared with the reference beams.
- When changing position and the number of couplers, there is a difference in the relatively low bearing capacity and ranges from 2 to 6%.
- The simulation method can predict and analyze the behavior of reinforced concrete beams with steel reinforced having the coupler.

For a complete evaluation of the behavior of reinforcement concrete structure with the steel reinforcement connected with couplers, it requires further subsequent studies in the different forms of structure and impacts load type.

## ACKNOWLEDGMENTS

The authors sincerely thank the University of Technology and Education Ho Chi Minh City to implement the title Code T2020-71TĐ.

## REFERENCES

- [1] B. MacKay, D. Schmidt and T. Rezanoff, Mechanical Connections of Reinforcing Bars, reported by ACI Committee 439 B, (1998).
- [2] Rasha T.S. Mabrouk, Ahmed Mounir, "Behavior of RC beams with tension lap splices confined with transverse reinforcement using different types of concrete under pure bending," Alexandria Engineering Journal, Vol.-57, Issue 3, P. 1727-1740, (2018).
- [3] Standards New Zealand, NZS 3101:2006: Concrete Structures Standard - Part 1: The Design of Concrete Structures sets out minimum requirements for the design of reinforced and pre-stressed concrete structures, (2006).
- [4] American Concrete Institute, ACI 318M-11: Building Code Requirements for Structural Concrete and Commentary, (2011).
- [5] Ahmed El-Azab, Hatem M. Mohamed, "Effect of tension lap splice on the behavior of high strength concrete (HSC) beams," HBRC Journal, Vol.-10, Issue 3, P. 287-297, (2014).
- [6] Vietnam Standards, TCVN 8163:2009 : "Steel for the reinforcement of concrete – Threaded coupler splice", (2009).
- [7] ABAQUS users' manual, Hibbitt Karlsson & Sorensen Inc, (2014).
- [8] [8] Le Anh Thang, Nguyen Tat Thanh, Hoang Trong Quang, "Validation of computational models of steel slag used as large particles in concrete beams", Internet: <http://thuvienso.hcmute.edu.vn>, (2016).
- [9] Swami P. S., Javheri S. B., Mittapalli D. L, Kore P. N., "Use of mechanical splices for reinforcing steel", Conference proceedings Issn No - 2394-3696, (2016)
- [10] Teuku Budi Aulia, Rinaldi, "Bending capacity analysis of high-strength reinforced concrete beams using environmentally friendly synthetic fiber composites", Procedia Engineering 125, P. 1121 – 1128, (2015)



# Optimum of Biodegradable Plate Production from Banana Trunk Waste by Taguchi Methods

Nhung Thi-Tuyet Hoang  
Faculty of Chemical and Food  
Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh city, Vietnam  
nhungtht@hcmute.edu.vn

Anh Thi-Kim Tran  
Faculty of Chemical and Food  
Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh city, Vietnam  
anhhtk@hcmute.edu.vn

Phan Thi Thu Thuy  
Faculty of Mechanical Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh city, Vietnam  
thuyptt@hcmute.edu.vn

**Abstract**—Single-use plastic waste disposal is a global problem affecting our planet not only with the pollution but also with the natural resource (fossil fuel) depletion. This study is aimed to provide a general waste recycling process from agriculture waste of banana fibre to produce biodegradable plates which are promised to replace single-use plastic products and reduce the usage of raw material for plastic production. With the characteristic of high strength fibre from the trunk, banana (*Musaceae*) can be applied for producing composite material or paper. These banana trunks were pretreated by cooking with sodium carbonate solution and then blended in a mixer for different timing to form a pulp. Additives (starch) with different percentage were added to the pulp after cooking to enhance the properties of plates. Finally, the pulp was formed by a plate machine to make products. To optimize the production process, the effects of sodium carbonate concentration, cooking time, percentage of starch, and thickness of product were evaluated using an L16 orthogonal array design by Taguchi method. The optimum conditions for suitable biodegradable plates production were determined as sodium carbonate concentration of 10%, cooking time of 45 minutes, 15% percentage of starch and weight of pulp, 45 g. The tensile stress and weight load of the optimum product were 8 N/mm<sup>2</sup> and 35 g/mm<sup>2</sup>, respectively. From these results, the product from the banana fibre can be believed to use as material for making biodegradable plates to mitigate the pollution problem in modern cities.

**Keywords**— *Banana fibre, Banana peel, Pulp, Taguchi method, Biodegradable plate*

## I. INTRODUCTION

Single-use plastic (SUP) which are accessible for food packaging is one of the chemically inert non-biodegradable plastic materials. Due to the difficulty of recycling this material, after its usage, SUP packaging is thrown away, threatening the quality of land, water, air and ocean. According to the official statistics of the Ministry of Natural Resources and Environment, Vietnam is among the top 4 countries wasting the most enormous amount of SUP all over the world, after China, Indonesia and the Philippines. The plastic consumption per capita increases from 3.8 kg to 41.3 kg after 28 years in Vietnam. Especially, the number of SUP, which are not recycled and must be buried, in the two largest cities of Vietnam, Ha Noi and Ho Chi Minh City rise to 80 tons each day. Plans are deploying to remove SUP in urban markets, convenience stores and supermarkets by 2021 and the whole country by 2025.

Many green products were studied to replace SUP, such as using banana leaves instead of plastic bags to wrap vegetables; or using paper packaging to wrap eggs; bagasse

containers; paper packing, etc. Products from the banana trunk (green packing materials) have been studied and deployed in developing countries. Banana (*Musa paradisiaca*, family *Musaceae*) is one of the major fruits in Vietnam with 100.000 ha and 1.4 million tons of banana per year. After harvesting, the banana trunk is usually used for feeding animal and making organic fertilizer. Banana trunk (BT) could be also used as raw material in the processing of biodegradable materials because of its high cellulose (50-60%), hemicelluloses (25-30%), pectin (3-5%), lignin (12-18%), water-soluble materials (2-3%), fat and wax (3-5%) and ash (1-1.5%) [1]; [2]. In recent years, many studies are focused on recycling banana trunk for pulping of banana fibre [3], nanocellulose fiber for a composite film [4], microcrystalline cellulose [5], fabric and apparel [6], eco-bag [7]. With the aforementioned studies, the banana trunk is suitable for the biodegradable product for food containers. To achieve this container application, the banana needs to be reinforced to obtain the ability of tensile stress as well as weight load. Several methods for the design and optimization of experiments were studied to minimum the number of tests. One of well-known methods is Taguchi approach. In Taguchi method, results of experiments was analyzed as a statistical method to investigate how different parameters affect the mean and variance of the synthesis or operation process [8].

In this work, the single-use plate from the banana trunk was produced with simple process by cooking banana trunk with sodium carbonate ( $\text{Na}_2\text{CO}_3$ ), then adding starch to the solution and plate printing by machine. The BT plate was produced on a machine which was developed by the HCMUTE innovator. The effect of operational parameters such as  $\text{Na}_2\text{CO}_3$  concentration, cooking time, starch concentration and weight of pulp on the mechanical properties of products (tensile stress, weight load) were determined and optimized by Taguchi method. Modelling the effects of products, identifying the influences of four factors on tensile stress and weight load and their interactions were conducted using general regression and analysis of variance (ANOVA).

## II. EXPERIMENTAL

### A. Preparation of banana peel plate

Banana trunk (BT) were collected from the local rural areas of Vietnam. BP was cut to 0.5 – 1.5 cm in length and dried under the sun. After that, BT was alkalized with  $\text{Na}_2\text{CO}_3$  and then cooked at 100°C. The pulp solution was filtered and cleaned with distilled water until neutral pH. Starch was added before the compressing process. The



process of pulp compressing was done in a metal mould with the dimension in Figure 2. The plate then was dried in the oven of 40-50°C for 12 hours. For waterproofing, the lotus leaves were compressed on the plate after washed and dried at 40°C in 15 - 60 minutes.

### B. Design of procedures of biodegradable plate

Four effective factors including  $\text{Na}_2\text{CO}_3$  concentration (A), cooking time (B), percentage of starch (C) and pulp weight (D) were chosen to analyze the optimum value for the producing process. Every factor was performed four levels (Table I). With four levels, the full factorial experimentation required  $4^4 = 256$  experiments. To save time and experiments, Taguchi approach was used in our present work to determine the orthogonal array for experimental design and signal to noise ratio (S/N) for quality [9, 10]. In Taguchi method, the necessary number of experiments for 4 factors with four levels was 16, which enough to study the main effect and correlations. Therefore, the Taguchi L16 orthogonal array (L16-OA) was carried out. All tests were repeated three times.

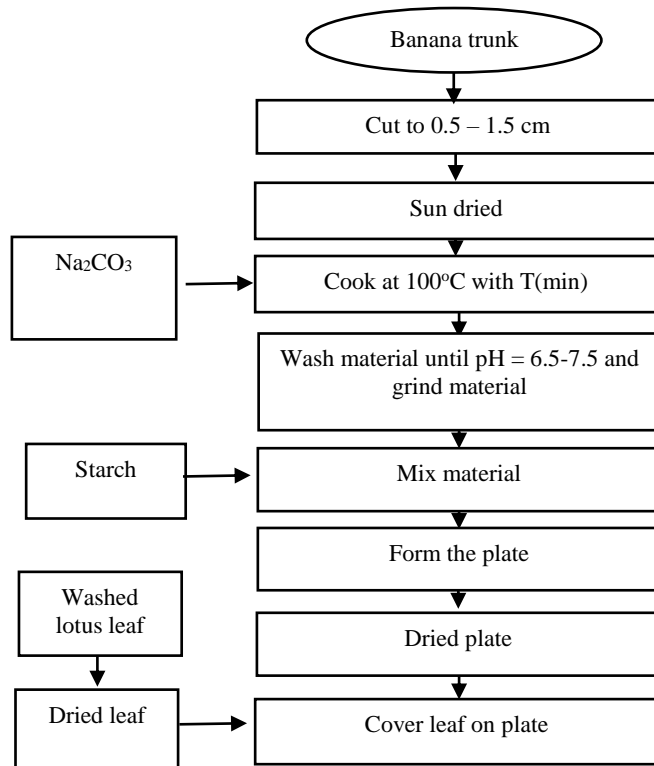


Fig. 1. Procedure for biodegradable plate from banana trunk

TABLE I. FACTORS AND THEIR LEVELS USED FOR THE DESIGN OF EXPERIMENTS BY TAGUCHI METHOD.

Independent factors	Level 1	Level 2	Level 3	Level 4
<b><math>\text{Na}_2\text{CO}_3</math> concentration (% w/v)</b>	5	10	15	20
<b>Cooking time, min</b>	15	30	45	60
<b>Starch, % w/v</b>	5	10	15	20
<b>Weight of pulp, g</b>	30	45	60	75

### C. Material characterization

#### Mechanical Properties

Testometric M500-50 machine was used for determining the mechanical properties of BT plate such as tensile strength and strain break. The tests were performed at room temperature with rectangular samples of dimensions 25 cm by 60 cm. The two ends of the sample were then kept with 30 mm of length between two mounts and a crosshead speed of 0.5 mm/min. Tensile strength and strain break were recorded from stress-strain curves. All tensile property measurements of the testing samples were performed on the averages of three reported values

#### Thickness and density

The thickness of BT plate was determined with a manual micrometer (Mitutoyo, Japan). The density was calculated as the relationship between weight and thickness. The reported value is the average of 5 determinations.

#### Weight load

The sample was put on the glass cup how to create free space under the sample. Heavy material was added on the surface of the sample for 1 minute. Determine the weight loading by increasing weight on the BT plate until it breaks. The average result of the three repeated times.

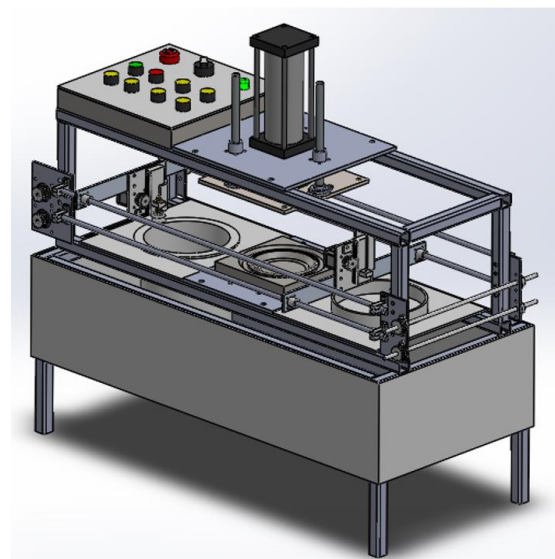
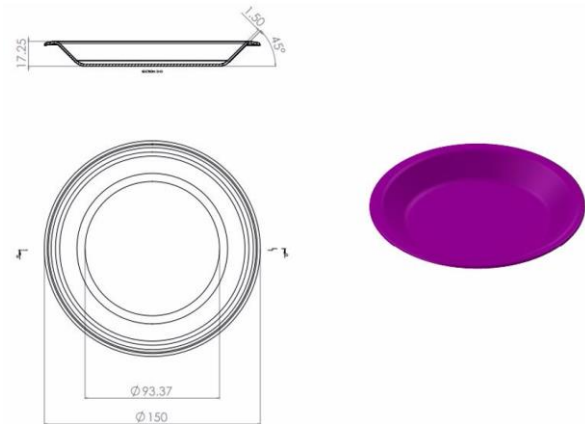


Fig. 2. Dimension of biodegradable plate and compressing machine

### III. RESULTS AND DISCUSSION

#### A. Taguchi analysis

The Taguchi method was used through Minitab 19 software to determine the optimum conditions and analyze which factor effects on processing BT plate. The tensile strength (N) and weight load of the BT plate were used as an index for evaluation of mechanical properties. The results of

the tensile stress and weight loading obtained from the tests are statistically analyzed using S/N ratios. The average S/N ratio of factors at different levels are summarized in Table II. Because of the higher tensile stress and weight load, the better is BT plate expected, the S/N ratios are analyzed based on 'larger is better' for optimum choice. It means that the maximum S/N of factors are corresponding to the maximum tensile strength and weight load.

TABLE II. EXPERIMENTAL PLAN AND RESULTS USING THE L4 (44) ORTHOGONAL ARRAY

Na <sub>2</sub> CO <sub>3</sub> (A), % w/v	Cooking time (B), min	Starch (C), % w/v	Weight of pulp (D), g	Tensile stress N/mm <sup>2</sup>	Weight load (g/cm <sup>2</sup> )	SN ratio
5	15	5	30	5.59	2.36	9.76
5	30	10	45	7.43	7.36	17.38
5	45	15	60	7.71	12.06	19.26
5	60	20	75	5.71	21.51	17.85
10	15	10	60	4.63	4.67	13.35
10	30	5	75	8.02	15.98	20.12
10	45	20	30	7.90	31.64	20.70
10	60	15	45	7.97	33.4	20.80
15	15	15	75	6.11	3	11.61
15	30	20	60	5.94	17.54	18.01
15	45	5	45	8.85	24.27	21.40
15	60	10	30	8.35	38.4	21.24
20	15	20	45	6.41	4.42	14.23
20	30	15	30	8.67	11.51	19.82
20	45	10	75	7.57	38.63	20.43
20	60	5	60	5.85	45.42	18.29

The S/N ratio considerably increases when cooking time increases from 15 to 45 min, which indicates that the quality of BT plate was improved with the rise in the cooking time; however, with more than 45 min, this improvement is not great. This can be explained that when longer cooking makes the polysaccharide degraded, therefore, induces the tensile stress. Similarly, the S/N ratio is the highest when the Na<sub>2</sub>CO<sub>3</sub> concentration increases to 10%, then decreases with a higher concentration of Na<sub>2</sub>CO<sub>3</sub>. With a lower degree of alkali degree, the reaction to remove lignin occurs weakly, however, with a higher degree, cellulosic fibres was degraded leading to the reduction in tensile stress [11]. For starch concentration and pulp weight, the S/N ratio slightly changes with the increase in level. With all above analysis, the optimum conditions to get the highest value for the average tensile stress and weight load are the concentration of Na<sub>2</sub>CO<sub>3</sub> level 2 (10%), cooking time level 3 (60 min), percentage of starch level 2(10%) and weight of pulp level 2 (45g).

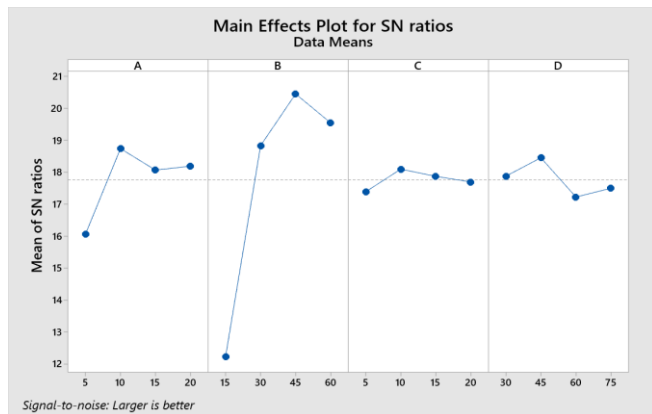


Fig. 3. Effect of factors on tensile stress and weight load of BT plate.

To discern the effective factors on the BT plate process, an analysis of Taguchi was presented in Table III. The most influential factor is cooking time (min) because of the highest variance (rank number 1). The second effective factor belongs to the concentration of Na<sub>2</sub>CO<sub>3</sub> (rank number 2) while the variation of the percentage of starch and the weight of pulp do not much change with four levels.

TABLE III. DETERMINATION OF THE SIGNIFICANT FACTORS

Level	Na <sub>2</sub> CO <sub>3</sub> , % w/v	Cooking time, min	Starch, % w/v	Weight of product, g
1	16.06	12.24	17.39	17.88
2	18.74	18.83	18.10	18.45
3	18.07	20.45	17.87	17.23
4	18.19	19.54	17.70	17.50
Delta	2.68	8.21	0.71	1.23
Rank	2	1	4	3

#### B. Mathematic modelling

A general regression for the estimated model was performed from the experimental results of tensile stress and weight load. The model of the process and the interaction among factors affecting the mean of tensile stress and weight load are shown in Equation (1) and (2):

$$\text{Tensile stress (N/mm}^2\text{)} = -0.28 + 0.3808 A + 0.5062 B - 0.1919 C - 0.0363 D - 0.003185 B^2 - 0.01393 AB + 0.00316 AD - 0.0012 BD \quad (1)$$

$$\text{Weight load} = 10.03 - 1.835 A + 0.331 B + 0.267 C - 0.319 D + 0.0327 AB + 0.0276 AD \quad (2)$$

Where A is the concentration of  $\text{Na}_2\text{CO}_3$  (w/v), B is the cooking time (min), C is the percentage of starch (% w/v) and D is the weight of pulp (g)

Analysis of variance (ANOVA) for tensile stress and weight load (Table IV and V) was done to find out which parameters significantly affect independent variables and the interaction between different parameters. High obtained F-values and correspondingly low  $p$ -values perform highly confidence of factor or regression. When the regression model has  $p$ -value  $< 0.05$ , the confidence of its statistical significance is higher 95%. In this study, both regression equations for tensile stress and weight load are generated with  $p$ -value  $< 0.05$ . It means that both models are high reliability, accuracy with small error. The fitness of model regression is also presented by the coefficient  $R^2$ , which is the ratio of regression variation to total variation. The more similar between predicted and experimental results, the closer to 1 is  $R^2$ . The  $R^2$  of the regression model for tensile stress and weight load are 94.12% and 95.32%, respectively. The importance of the parameters, interaction and square terms were analyzed by Minitab 19, and the insignificant terms of  $p$ -value which are much higher than 0.05 are eliminated manually.

To analyze the effect of tensile stress, the P-value of concentration of  $\text{Na}_2\text{CO}_3$  (A), cooking time (B), percentage of starch (C), AB and BD are lower than 0.05. It means the high significance of factor A, B, C and high interactions between factors (A and B, D and D). Presence of interaction terms between A via B and B via D (i.e., the concentration of  $\text{Na}_2\text{CO}_3$  via cooking time; cooking time via weight of pulp) in the regression model of tensile stress shows that these parameters were not completely independent.

To analyze the effect of weight load, the only  $p$ -value of concentration of  $\text{Na}_2\text{CO}_3$  (A) is lower than 0.05, corresponding to its high significance. Other factors do not much affect the weight load. Besides, there is no evidence for the interaction between elements.

TABLE IV. ANALYSIS OF VARIANCE FOR REGRESSION MODEL OBTAINED FROM THE EXPERIMENTAL TENSILE STRESS OF BT PLATE

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	23.2063	2.9008	14.00	0.001
A	1	4.1080	4.1080	19.82	0.003
B	1	15.7159	15.7159	75.83	0.000
C	1	5.0567	5.0567	24.40	0.002
D	1	0.3066	0.3066	1.48	0.263
B*B	1	8.2191	8.2191	39.66	0.000
A*B	1	4.7097	4.7097	22.73	0.002
A*D	1	0.6368	0.6368	3.07	0.123
B*D	1	1.6527	1.6527	7.97	0.026
Error	7	1.4507	0.2072		
Total	15	24.6570			

TABLE V. ANALYSIS OF VARIANCE FOR REGRESSION MODEL OBTAINED FROM THE EXPERIMENTAL WEIGHT LOAD OF BT PLATE

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	2915.50	485.917	30.57	0.000
A	1	95.54	95.538	6.01	0.037
B	1	15.39	15.392	0.97	0.351
C	1	9.82	9.816	0.62	0.452
D	1	32.27	32.274	2.03	0.188
A*B	1	25.90	25.904	1.63	0.234
A*D	1	48.39	48.393	3.04	0.115
Error	9	143.06	15.896		
Total	15	3058.56			

The significant and non-significant terms of the BT plate process was also represented by Pareto chart. In standardized PARETO chart, the magnitude and the importance of factors as well as the interaction are described by bars [11]: the more extended bar, the more critical factor for BT plate process. The length of the bar over the vertical line, which is drawn at the location of the 0.05 critical values for Student's  $t$  indicate effects that are statistically significant at the 5% significance level. As can be seen in Fig.4, the cooking time (B) and concentration of  $\text{Na}_2\text{CO}_3$  (B) are the most significant parameter (highest level of standardized effect) for tensile stress and weight load, respectively.

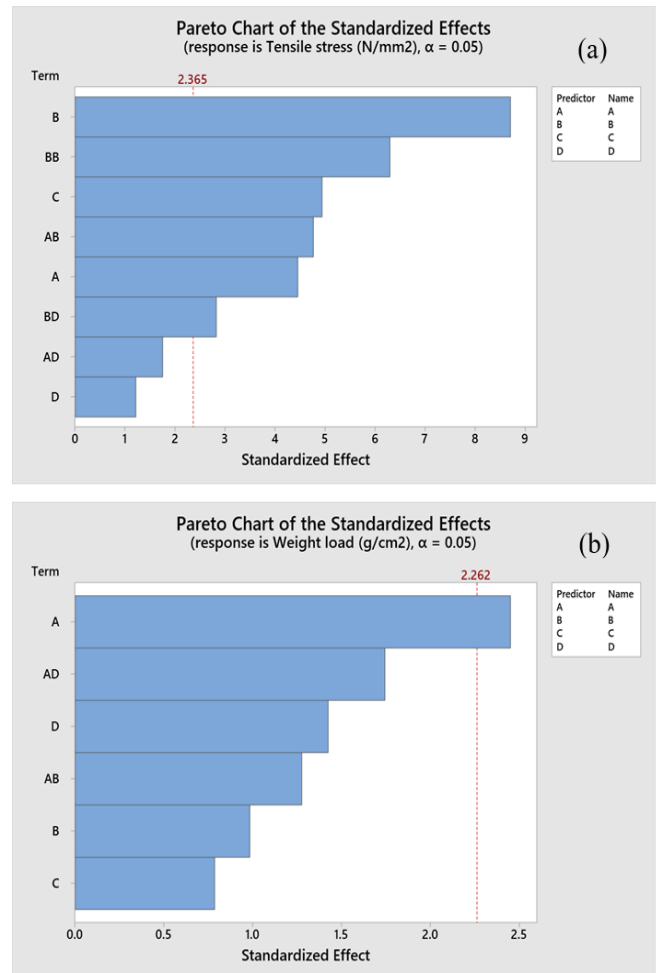


Fig. 4. Pareto chart of standardized Effects for tensile stress (a) and weight load (b)

### C. Material characterization



Fig. 5. BT plate without and with lotus leaf

After determining the optimal conditions for the producing BT plate by analyzing the Taguchi method, the product was processed to evaluate the characteristics of BT plate. The BT plate has a thickness from 0.6 – 0.8 cm, with a density of 0.005 g/cm<sup>3</sup>. The tensile stress was  $7.77 \pm 1.2$  N/mm<sup>2</sup>, and weight load was 33.4 g/cm<sup>2</sup>. Fige. 5 shows the obtained single-use dish from banana fibre without and with lotus leaf, which can be used for food. This can be a replacement for plastic bags to help to reduce plastic consumption and protect the environment.

#### IV. CONCLUSION

The main objective of this study was to study the feasibility of a banana trunk as a potential source of single-use materials. By the optimization process using an L16 orthogonal array design with Taguchi method, it can be concluded that: (1) the optimum conditions for suitable biodegradable plates production were determined as sodium carbonate concentration of 10%, the cooking time of 45 minutes, 15% percentage of starch and product thickness of 0.7 mm; (2) the tensile stress and weight load of the optimum product were 8 N/mm<sup>2</sup> and 35 g/mm<sup>2</sup>, respectively; (3) the model of the process, as well as the interaction among factors which affect the mean of tensile stress and weight load, was established with  $p$ -value  $\ll 0.05$  and  $R^2$  higher than 94%. With these obtained results, this banana trunk after cooking conditions can be used as raw materials making single-use dish, which can help reducing plastic usage demand to protect the environment.

#### REFERENCES

- [1] M. Omotoso and B. Ogunsile, "Fibre and Chemical Properties of Some Nigerian Grown Musa Species for Pulp Production," *Asian Journal of Materials Science*, vol. 1, pp. 14-21, 01/01 2009.
- [2] S. Mukhopadhyay, R. Fanguiero, Y. Arpaç, and Ü. Sentürk, "Banana Fibers – Variability and Fracture Behaviour," *Cellulose*, vol. 31, p. 3.61, 06/01 2008.
- [3] M. Abd Rahman, A. Jasani, and M. Ibrahim, "Flexural Strength of Banana Fibre Reinforced Epoxy Composites Produced through Vacuum Infusion and Hand Lay-Up Techniques - A Comparative Study," *International Journal of Engineering Materials and Manufacture*, vol. 2, p. 31, 07/01 2017.
- [4] B. H. Patel and P. V. Joshi, "Banana Nanocellulose Fiber/PVOH Composite Film as Soluble Packaging Material: Preparation and Characterization," *Journal of Packaging Technology and Research*, vol. 4, no. 1, pp. 95-101, 2020/03/01 2020.
- [5] N. Shanmugam, R. D. Nagarkar, and M. Kurhade, "Microcrystalline cellulose powder from banana pseudostem fibres using bio-chemical route," *Indian Journal of Natural Products and Resources*, vol. 6, pp. 42-50, 03/01 2015.
- [6] M. Maleque, B. Yousif, and S. Sapuan, "Mechanical properties study of pseudo-stem banana fiber reinforced epoxy composite," *Arabian Journal for Science and Engineering*, vol. 32, 10/01 2007.
- [7] D. Mohapatra, S. Mishra, and N. Sutar, "Banana and its biproduct utilization," *J. Sci. Ind. Res.*, vol. 69, pp. 232-329, 01/01 2010.
- [8] A. Panda and R. K. Singh, "Optimization of process parameters in the catalytic degradation of polypropylene to liquid fuel by Taguchi method," *Adv Chem Eng Res (ACER)*, vol. 2, pp. 106-112, 01/01 2013.
- [9] G. Taguchi, *The System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs*. Quality Resources, 1987.
- [10] P. J. Ross, *Taguchi Techniques for Quality Engineering*. New York: McGraw-Hill, 1988.
- [11] N. Cordeiro, N. Belgacem, I. C. Torres, and J. V. C. P. Moura, "Chemical composition and pulping of banana pseudo-stems," *Industrial Crops and Products*, vol. 19, pp. 147-154, 03/01 2004.

# Wheelchair Navigation System using EEG Signal and 2D Map for Disabled and Elderly People

Ba-Viet Ngo

*Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education,  
Ho Chi Minh City, Vietnam  
vietnb@hcmute.edu.vn*

Thanh-Hai Nguyen

*Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education,  
Ho Chi Minh City, Vietnam  
nthai@hcmute.edu.vn*

Van-Thuyen Ngo

*Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education,  
Ho Chi Minh City, Vietnam  
thuyen.ngo@hcmute.edu.vn*

Dang-Khoa Tran

*Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education,  
Ho Chi Minh City, Vietnam  
tdkhoa@hcmute.edu.vn*

Truong-Duy Nguyen

*Faculty of Electrical and Electronics Engineering  
Ho Chi Minh City University of Technology and Education,  
Ho Chi Minh City, Vietnam  
duynt@hcmute.edu.vn*

**Abstract**— Disabled and elderly people may be very difficult to control an electrical wheelchair by using a joystick in environments with tight spaces, narrow pathways, and obstacles. This paper proposes a wheelchair system with hands-free control to assist users to overcome this difficulty. In particular, a combination of EEG signals and a 2-dimensional (2D) map is applied in this wheelchair system. Firstly, a human brain-computer communication system for interfacing with human facial gestures to execute commands using the EEG signals is designed. The EEG signals are processed using filtering, scaling, and then features are extracted to connect with the input of one neural network for classifying the EEG signals corresponding to control commands. Secondly, environmental information with obstacle and free spaces are collected for building a 2D map using an RGB-D camera system. Therefore, user inputs are used to a navigation plan along with the 2D safety map for a navigation wheelchair system with collision-free control during movement. Experimental results demonstrate to illustrate the feasibility and effectiveness of the proposed method.

**Keywords**— Smart wheelchair, BCI system, Shared control, EEG signal, RGB-D camera system, 3D depth map, 2D map, Obstacle avoidance

## I. INTRODUCTION

Electric wheelchairs are equipped with technologies that are being developed in recent years [1]. People with severe disabilities are the main users of the electric wheelchair system for independent mobility to desired destinations [2]. More often, they will find the conventional wheelchair considerably difficult to control by using their legs or arms. As a result, more researches are focusing on using biological signals generated from movements or functions of anatomical units of the human head as variable inputs for controllers. Those units include the head, eyes (oculography), muscles (myography) or even the brain (electroencephalogram - EEG) [3, 4]. Voice-activated navigation [5] requires a quiet environment and may not be good for use in busy and noisy environments. For brain waves [6], it has recently become a subject of interest in mechanical control. To do so, the EEG signal samples need to be classified and grouped into intended actions. The eye input provides good information such as the direction of the head and the eye [7] to manipulate. While this technique might seem like a good candidate, it's hard to

distinguish between the act of driving a wheelchair or simply looking around. The best solution is to use multiple inputs from the user [8, 9], in such a way that many possible user signals are analyzed before issuing the desired command. Using this strategy, they can designate each control task to correspond to different user inputs. Hence, this will create less burden on the user than relying on a single input.

Most recently, research [10] has shown that most patients with mobility impairment cannot control a wheelchair to avoid obstacles. These clinical findings provide insight into the importance of creating a computer-controlled platform to assist users by reducing their workload and increasing safety. In this framework, user input along with environmental information will be analyzed seamlessly to perform necessary support tasks. Supports can be categorized into three main levels; general control, semi-automatic control, and autonomous control. General controls allow the user to control the wheelchair completely. Computers only step in when necessary, for example in case the wheelchair goes through doors or avoids obstacles [11, 12]. Autonomous control allows a wheelchair to automatically move on the way to a user-selected final destination [13, 14]. Due to this nature, shared control is more suitable for users who can provide continuous input using the joystick [15]. Autonomous control is more suitable for users who cannot deliver low-end orders, people get tired easily, and people with visual impairments [16].

Since people with severe motor disabilities can still command their head even with a limited amount of movement, semi-autonomous control is a proper assistive choice. In this control mode, the computer performs short-term route planning, and the users only intervene when they wish to deviate from the plan [17, 18]. This means that, when the wheelchair operates in this mode, users can relax while the computer is completed task. Unlike the autonomous control, the semi-autonomous control does not need an actual map of the environment; only a local safety map based on sensors scanning is needed. As a result, it can give freedom to the users to move in new environments.

In this paper, we propose a wheelchair navigation system with a hybrid control of EEG signal and 2D map to aid the mobility of people with severe motor impairment. Since the user can issue a limited control command within a short period, the computer will take over the responsibilities for



navigating and avoiding any possible obstacles. In this situation, the user is only responsible for heading the wheelchair into the desired directions. With this method, critical and dangerous situations can be effectively overcome, while the user can still feel comfortable and safe while controlling the wheelchair.

## II. METHODS

### A. BCI based control system

In this system, an implementation of a brain-computer interface (BCI) method without a joystick is proposed. This method can minimize users' workload by controlling the wheelchair using facial movements.

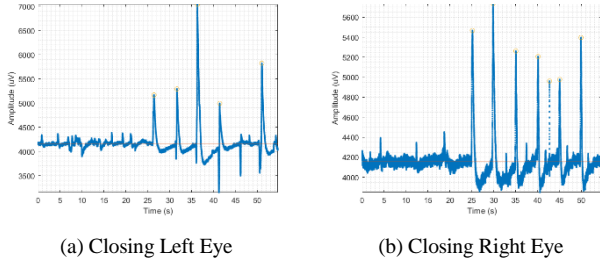


Fig. 1. EEG waveform with eye activities

Fig. 1 shows the eye activities, in which the movement of the left and right eyes corresponding to AF3 and AF4 channels, respectively. In addition, the BCI control system consists of an Emotiv headset connected to a computer and electrodes (electrochemical sensors) pick up facial movements to produce signals transferred to the computer. The signals are processed to produce features connected to a neural network for classifying. Results sent to a microcontroller are forward, turn left, turn right and stop commands for wheelchair control. The wheelchair is often set up at low speed for safety of user.

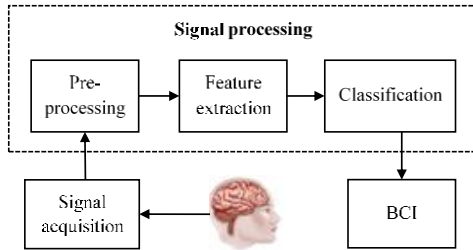


Fig. 2. General structure of BCI

In this research, the BCI-based control wheelchair system is composed of five main units: signal acquisition, signal preprocessing, feature extraction, classification, and control of motors as represented in Fig. 2. In the signal acquisition, EEG signals are captured using an Emotiv EPOC headset with 14-channels, in which its features are adequate for the useful BCI (resolution and bandwidth). In addition, the control wheelchair system uses upper face gestures for actuation commands the most reliable signals corresponding to Emotiv sensors installed in the frontal cortex. Thus, the EEG input signals are sent to the signal preprocessing unit for filtering and scaling to extract features before sending them to the classification system. After classified, the outputs corresponding to control instructions are sent to the motors for control of the wheelchair.

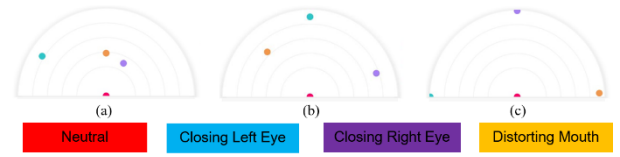
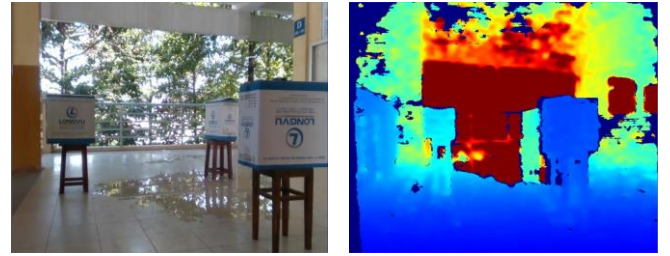


Fig. 3. Description of the training results of four movements from the EEG signal

Fig. 3 depicts the EmotivBCI software interface for training four types of facial movements corresponding to four different colors (red, blue, purple, and yellow). In Fig. 3a, four dots are to represent different types of error movement, in which training failure was due to the user's inability to follow the training instructions correctly. The training results could be improved as shown in Fig. 3b and the best classification results are shown in Fig. 3c, particularly the dots are spaced far apart on the semi-circle. This is a simple and efficient solution to evaluate the training process, helping the trainee to see the training results in the most intuitive and quickest way as well as possible.

### B. 3D-to-2D mapping process



(a) Environment with three obstacles (b) 3D depth map of the environment

Fig. 4. Real environment and 3D depth map

A 2D map provides information about obstacles and free space for making plans for controlling an electric wheelchair. The 2D map was converted from a 3D depth map using the RGB-D camera system as shown in Fig. 4 [19, 20]. Firstly, the 3D depth map removed the points with height elements that are higher than the safety height of the wheelchair, in which it is only considered points in front of the wheelchair. Secondly, the remaining points were re-arranged into columns along the y-axis and the minimum depth in the z-axis on the  $i$ th column ( $Z_{imin}$ ) was chosen along with the width value of those points in the x-axis. If a column does not have pixels in the field of view of the camera, the column position is processed as a blank and the conversion of 3D map to 2D one is calculated as follows:

$$Z_{imin} = \min(Z_{ij}) \quad , j = \overline{0, n} \quad (1)$$

in which  $Z_{imin}$  is selected corresponding to  $Y_{jmin}$  depending on the height of the wheelchair. The 2D map ( $X_i, Z_{imin}$ ) will depend on  $Y_{jmin}$  value.

The smallest depth values,  $Z_{imin}$ , were selected from the collection of the minimum depth values,  $Z_{imin}$ , based on the following equations:

$$Z_{imin} = \min(Z_{imin}) \quad , i = \overline{0, m} \quad (2)$$

The width of the space  $a_v$  ( $v = 1, 2, \dots$ ) in the 2D map ( $X_i, Z_{imin}$ ) is expressed as follows:

$$a_v = |X_{k1} - X_{k2}| \quad (3)$$

with coefficients  $k_1$  and  $k_2$  are the first and last elements on the X-axis of the free space  $v^{th}$ , at column of depth  $Z \geq Z_{min}$ .

### C. EEG and 2D map data fusion systems

Many methods with data fusion such as the technique based on EEG signals and gaze tracking via a webcam were applied for wheelchairs [21]. Therefore, the control of an electric-powered wheelchair was a hybrid and self-configured fusion technique between EEG signals, head movement, and eye-tracking [22]. In our study, we proposed an EEG and 2D MAP data fusion system, in which there are a hybrid control and auto-control during electric wheelchair movement as shown in Fig. 5.

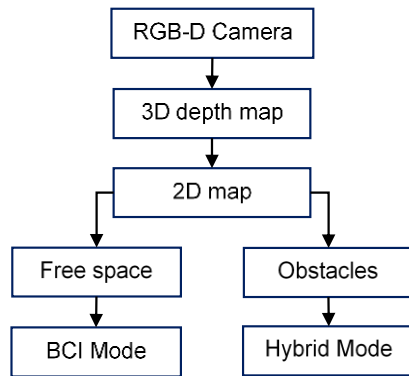


Fig. 5. Block diagram of two controls in the wheelchair navigation system

When the wheelchair moved within the setup environment, the 3D information of the environment will be converted into a 2D map. With this re-created 2D map, the system will identify the safe space in front of the wheelchair, which is identified based on the distance from the wheelchair to obstacles greater than 0.6m. If the frontal view is a safe space, the system will give users priority to control themselves via BCI. With this mode, users will use facial movements to control the wheelchair through the BCI protocol with the set of control command  $C = \{Forward, Right, Left, Stop\}$ .

In a situation where the wheelchair enters an unsafe area, the hybrid mode is activated, in which both BCI and camera control works together to form a decision. In particularly, the system will make a final control decision based on the consensus of both methods in order to avoid control errors that may lead to a collision. The system conducts sampling of control decisions for each method and calculates the probability for each control command in the control command set. Then, each method will select the commands with the highest probability to make a control decision. To formulate, let consider:

- $d_1$ : movement decision from the BCI control
- $d_2$ : movement decision from the control by the camera
- $d_{fusion}$ : the final selected command

The probability for each control command  $C_{ij}$  ( $i = \overline{1, 2}$  and  $j = \overline{1, 4}$ ), with the number of samples collected is  $n$ , is written as follows:

$$P(C_{ij}) = \frac{n(C_{ij})}{n} \quad (4)$$

The decision to select a control command for each type of input is calculated as follows:

$$d_i = \arg \max_{C_{ij}} (P(C_{ij})) \quad (5)$$

Finally, the  $d_{fusion}$  is calculated using the rules of consensus as follows:

$$d_{fusion} = \begin{cases} C_j & \text{if } d_1 = d_2 \\ Stop & \text{otherwise} \end{cases} \quad (6)$$

## III. RESULTS AND DISCUSSION

In this section, we analyze the system performance in real operating environments. The wheelchair was designed to move with a speed of 3 km/h in the indoor environment. The first experiment was conducted to evaluate the system's ability for responding to the user command via the designed BCI system. A user drove the wheelchair into predefined locations in the cluttered lab environment by the BCI control and the hybrid control. Based on these experiments, we describe an evaluation of the rehabilitation device. We focus on two different studies: a performance study of the intelligent wheelchair and a variability study among trials and subjects.

### A. EEG Training Results

After processing and feature extraction, signals were trained to classify through EmotivBCI software. During the training, the user performed facial movements such as closing the left eye, closing the right eye, glancing left, glancing right, distorting mouth and the relaxed state, without performing any activities. The training process is carried out twice and then the evaluation of the training results will be conducted. Training is considered good if the accuracy of the trained action is over 80%.

TABLE I. ACCURACY OF FACIAL MOVEMENTS OF FIVE USERS

Movement type	User 1	User 2	User 3	User 4	User 5
Glancing Left	70%	10%	20%	20%	40%
Glancing Right	60%	30%	70%	60%	10%
Distorting mouth	100%	90%	70%	90%	80%
Closing Left Eye	80%	100%	60%	100%	80%
Closing Right Eye	90%	80%	90%	90%	100%

Table 1 describes the experimental results when identifying five types of movements made by five subjects. Each person will perform ten movements for each type, each movement is measured in two seconds on average. The results showed high accuracy with an approximation of 95% for each of the three types: distorted mouth, left or right eyes closed. Meanwhile, the two types of eye movements to the left and

the right to glance have errors higher than 70%. The reason for this result is that the aiming movements are stronger and more clear than the squint movements which often make it difficult for the user to try to perform.

### B. Experimental Results

An electrical wheelchair was installed with an RGB-D camera system and other equipment as shown in Fig. 6. Information about the surrounding environment obtained from the camera system will be processed by a computer and then transferred to the motor system of the wheelchair for motion control.

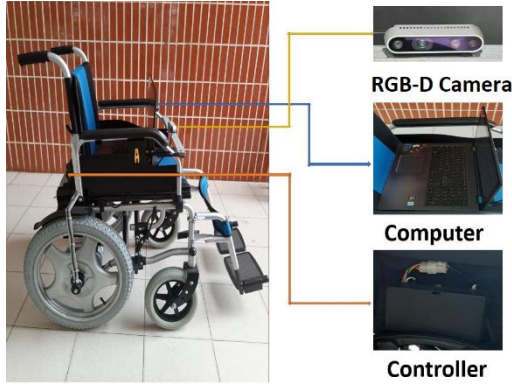
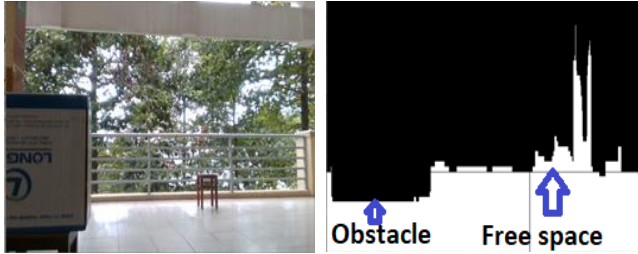


Fig. 6. The wheelchair navigation system installed with devices

After the creation of the 2D map, it showed the obstacles and free space in front of the camera, as seen in Fig. 7.  $Z_{min}$  depth of the obstacle area will be used to establish a safe distance for wheelchairs. At the same time, the width of the free space was compared against the width of the wheelchair. As a result, the wheelchair will only be allowed to move through any free space with the largest width.



(a) Real environment with a box and one chair

(b) 2D map corresponding to the real environment

Fig. 7. Environmental 2D map

Experimental results are demonstrated in Fig. 8, in which Fig. 8a and Fig. 8b showed that the wheelchair saw the free space on the way and then it passed through the free space. After that, the wheelchair encountered obstacles in front and it turned right for collision avoidance as shown in Fig. 18c and Fig. 18d. Continuously, in Fig. 8e and Fig. 8f it passed through the left free space and then moved forward to reach the target.

Fig. 9 illustrates a view of the path in the three trials, in which all measurements are in meters. The shaded objects represent static obstacles. The green dash line is the standard example of a reference path. These trials show that wheelchairs applied to the hybrid control method have the same trajectory. The wheelchair path is close to the actual path in the environment, showing that the method used is appropriate. The travel time for the three cases is shown in Fig. 10. The average travel time from the starting point to the

stopping point is 142s with a distance of 6.2m including obstacles, while that time of 325s with a distance of 12m in the study [18].



Fig. 8. Wheelchair moving in constrained environment

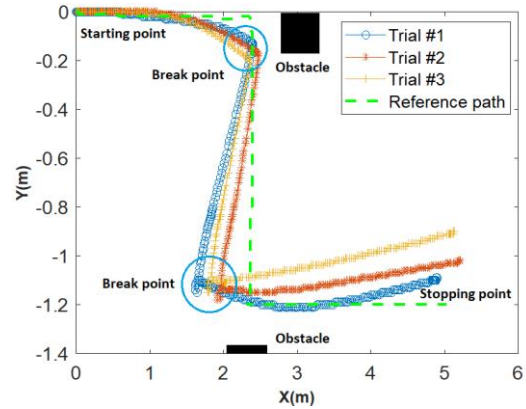


Fig. 9. Wheelchair trajectory in 3 trials compared to the green dotted reference path

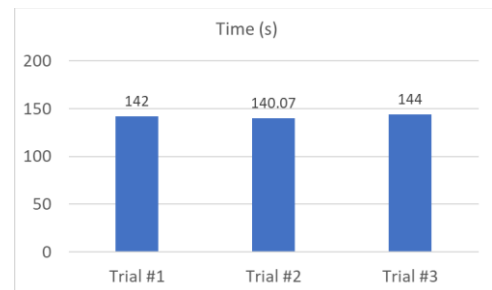


Fig. 10. Representation of movement time of the wheelchair in three trials

The experiment in Fig. 11 is to compare the wheelchair trajectory in two cases: using hybrid control and self-control using the BCI. The results showed that the wheelchair's path in the case of the hybrid control was safer than the self-control



one. In full self-control situations using the BCI, the possibility of a collision is possible (blue circle), endangering the user. The travel time of the wheelchair applied to the hybrid control is more than that of the BCI control.

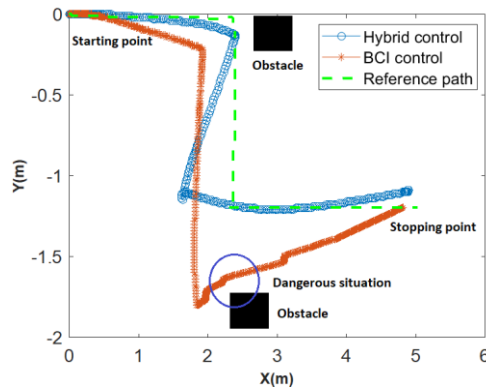


Fig. 11. Wheelchair trajectory in two control methods

#### IV. CONCLUSION

In this paper, a wheelchair navigation system was designed to assist elderly and disabled people in daily activity. The wheelchair could be controlled by a hybrid control method based on a fusion system with EEG signals and a 2D safety map. With the safety map, the navigation wheelchair could avoid collisions in both modes during movement in environments with obstacles in narrow spaces. With the proposed navigation wheelchair system, experimental results in actual cluttered environments showed that the wheelchair movement without obstacle collision is effective and safe for elderly and disabled people. In future work, we will evaluate of the reliability the proposed system through more participants. Moreover, the navigation wheelchair system needs to be developed by measuring the cognitive complexity of users with the extra workload during navigation.

#### ACKNOWLEDGMENT

We would like to thank the Faculty of Electrical and Electronics Engineering - HCMC University of Technology and Education, Viet Nam. Finally, an honorable mention goes to our volunteers and colleagues for their supports on us in completing this project.

#### REFERENCES

- [1] Al-Qaysi ZT, Zaidan BB, Zaidan AA, Suzani MS, "A review of disability EEG based wheelchair control system: Coherent taxonomy, open challenges and recommendations," *Comput Methods Programs Biomed*, vol. 164, pp. 221 - 237, 2018
- [2] Simpson RC, Lo Presti EF, Cooper RA, "How many people would benefit from a smart wheelchair?," *J of Rehab. Research and Development*, vol. 45, pp. 53-72, 2008
- [3] Ngo Ba Viet, Nguyen Thanh Hai and Ngo Van Thuyen, "Hands-free control of an electric wheelchair using face behaviors," *International Conference on System Science and Engineering (ICSSE)*, pp. 29-33, 2017
- [4] Chanlit Noiruxsar and Pranchalee Samanpiboon, "Face Orientation Recognition for Electric Wheelchair Control", *Journal of Automation and Control Engineering*, vol. 2, pp. 402-405, 2014

- [5] Nathalia Peixoto, Hossein Ghaffari Nik, Hamid Charkhkar, "Voice controlled wheelchairs: Fine control by humming," *Computer Methods and Programs in Biomedicine*, vol. 112, pp. 156-165, 2013
- [6] Jingsheng Tang, Yadong Liu, Dewen Hu and ZongTan Zhou, "Towards BCI-actuated smart wheelchair system," *BioMed Eng OnLine*, vol. 17, pp. 1-22, 2018
- [7] N. Wanluk, S. Visitsattapongse, A. Juhong and C. Pintavirooj, "Smart wheelchair based on eye tracking," *9th Biomedical Engineering International Conference (BMEiCON)*, pp. 1-4, 2016
- [8] Reis L.P., Braga R.A.M., Sousa M., Moreira A.P., "IntellWheels MMI: A Flexible Interface for an Intelligent Wheelchair," In: Baltes J., Lagoudakis M.G., Naruse T., Ghidary S.S. (eds) *RoboCup 2009: Robot Soccer World Cup XIII. RoboCup 2009*, *Lecture Notes in Computer Science*, vol. 5949, pp. 296-307, 2010
- [9] Sharmila Ashok, "High-level hands-free control of wheelchair – a review", *Journal of Medical Engineering & Technology*, vol. 41, pp. 46-64, 2017
- [10] Wan LM, Tam E, "Power wheelchair assessment and training for people with motor impairment", *12th Intl. Conf. on Mobility and Transport for Elderly and Disabled Person*, 2010
- [11] Amiel Hartman, Vidya K. Nandikolla, "Human-Machine Interface for a Smart Wheelchair", *Journal of Robotics*, vol. 2019, pp. 1-11, 2019
- [12] N. B. Viet, N. T. Hai and N. V. Hung, "Tracking landmarks for control of an electric wheelchair using a stereoscopic camera system," *International Conference on Advanced Technologies for Communications (ATC 2013)*, pp. 339-344, 2013
- [13] Tom Williams and Matthias Scheutz, "The state-of-the-art in autonomous wheelchairs controlled through natural language: A survey", *Robotics and Autonomous Systems*, vol. 96, pp. 171-183, 2017
- [14] K. Salhi, A. M. Alimi, P. Gorce and M. M. Ben Khelifa, "Navigation assistance to disabled persons with powered wheelchairs using tracking system and cloud computing technologies," *IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, Grenoble, pp. 1-6, 2016
- [15] Li, Zhijun et al, "Human Cooperative Wheelchair With Brain-Machine Interaction Based on Shared Control Strategy," *IEEE/ASME Transactions on Mechatronics*, vol. 22, pp.185-195, 2017.
- [16] Carlson T, Demiris Y, "Increasing robotic wheelchair safety with collaborative control: Evidence from secondary task Experiments," In *Proc. Of Intl. Conf. on Robotics and Automation*, pp. 5582-5587, 2010
- [17] Lele Xi and Motoki Shino, "Shared Control of an Electric Wheelchair Considering Physical Functions and Driving Motivation", *International Journal of Environmental Research and Public Health*, vol. 17, pp. 5502-5520, 2020
- [18] J. Duan, Z. Li, C. Yang and P. Xu, "Shared control of a brain-actuated intelligent wheelchair," *Proceeding of the 11th World Congress on Intelligent Control and Automation*, pp. 341-346, 2014
- [19] N T Hai and N T Hung, "A Bayesian Recursive Algorithm for Freespace Estimation Using a Stereoscopic Camera System in an Autonomous Wheelchair," *American J. of Biomedical Eng*, vol. 1, pp. 44-54, 2011.
- [20] N T Hai and V T Kiet, "Freespace Estimation in an Autonomous Wheelchair Using a Stereoscopic Cameras System," *The 32rd IEEE Annual International Conference on EMBS*, 2010.
- [21] F. Ben Taher, N. Ben Amor and M. Jallouli, "A multimodal wheelchair control system based on EEG signals and Eye tracking fusion," *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, Madrid, 2015, pp. 1-8.
- [22] Fatma Ben Taher, Nader Ben Amor, Mohamed Jallouli, "A self configured and hybrid fusion approach for an electric wheelchair control", *Intelligent Systems (IS) 2016 IEEE 8th International Conference on*, pp. 729-734, 2016.

# Experimental Study of the Strain Localization in a Rock Analogue Material at Brittle-Ductile Transition

Thi-Phuong-Huyen Tran

Department of Civil Engineering,  
UTE University of Technology and  
Education – The University of Danang  
Da Nang city, Viet Nam  
huyen.tran158@gmail.com  
tphuyen@ute.udn.vn  
<https://orcid.org/0000-0002-9164-9109>

Sy-Hung Nguyen

Department of Transport Engineering,  
Faculty of Civil Engineering  
HCMC University of Technology and  
Education  
Ho Chi Minh city, Viet Nam  
sihung.nguyen@hcmute.edu.vn

Stéphane Bouissou

Université Côte d'Azur, CNRS,  
Observatoire de la Côte d'Azur,  
Géozur  
Valbonne, France  
bouissou@geoazur.unice.fr

**Abstract**—Many studies on strain localization of geomaterials have been carried out over the past few decades. Theoretical analysis and numerical models show that the dilatancy property strongly affects the onset of the deformation localization bands. The evolution of deformation bands, however, remains poorly known. In this work, we analyzed the initiation and evolution of strain localization in a Granular Rock Analogue Material (GRAM1) performed under axisymmetric compression at brittle-ductile transition by using Digital Image Correlation (DIC) technique. DIC provides a full-field measurement of displacement and thus allows the analysis of non-homogeneous deformation. The onset of band/fracture localization is well detected and located on the stress-strain curve. During the deformation process, the deformation bands as well as sets of sub-parallel strands which contribute the final thickness of the main shear band form sequentially. The analyses of DIC data and post-mortem sample observation enhance the understanding of the strain localization phenomena at brittle-ductile transition.

**Keywords**—Deformation bands, brittle-ductile transition, digital image correlation.

## I. INTRODUCTION

Rock failure is generally classified in three regimes: brittle, ductile and transitional brittle-ductile regimes. A large amount of experimental tests of rocks performed under axisymmetric compression, covering the whole range of material behavior from brittle to brittle-ductile transition, showed the important role of porosity change induced by the deviatoric stresses over the failures modes. In the brittle regime, the deviatoric stress can cause dilatancy which ultimately leads to failure by shear dilatancy band localization and brittle faulting. During cataclastic flow, the deviatoric stress provides significant contribution on the compaction strain which induces “shear-enhanced compaction” phenomenon [1]. The reduction of porosity leads the rock to become harder, thus inhibiting the development of shear localization. Byerlee [2] studied the frictional characteristics of rocks and found that the brittle-ductile transition pressure  $\sigma_{bdt}$  is the pressure at which the stress required to form a fault surface is equal to the stress required to cause sliding on the fault. At  $\sigma_m < \sigma_{bdt}$  ( $\sigma_m$  – mean stress), the failure envelope is rather linear, the failure corresponds to an abrupt stress drop for axisymmetric compression conditions. At  $\sigma_m > \sigma_{bdt}$ , its slope progressively reduces to zero and then to the negative values, there is a smooth stress reduction with an axial strain

increase [1]. The transition between two regimes, which respectively lead to a localized strain with dilatancy inside a shear band and a quasi-homogeneous strain with compactancy, is still not well understood. In the brittle-ductile transition, a set of conjugated shear bands is typically formed, and becomes more and more numerous and closely spaced [3]. The evolution of this process, including the moment and place of the onset of the first band and then the others, however remains poorly known. The detection of the onset of deformation banding is crucial for the theoretical analysis of the underlying mechanism, which is believed to be a constitutive instability resulting from the deformation bifurcation [4]. One of the major criteria of applicability of the theory is a correspondence the critical hardening modulus predicted theoretically to that measured in the experiments. To make a correct measurement, one has to precisely define the onset of band localization and locate it on the stress-strain curve.

In conventional triaxial compression tests, the sample is within the steel pressure cell, which makes it impossible to measure directly the deformation field unless using heavy and expensive X-ray tomography techniques that can be applied in certain cases [5], [6]. In most cases only the postmortem sample observation is possible. Over the past few decades, extensive experimental investigations on strain localization of geomaterials have been carried out. In such tests, full-field measurements are being used to study the strain localization phenomenon which provides information concerning deformation localization pattern, onset of band/fracture localization and its propagation to the ultimate state. One of the most widespread full-field measurement techniques is the method of displacement field measurement using Digital Image Correlation (DIC). DIC technique provides qualitative and quantitative descriptions of local responses, which allow evaluating the strain distribution and localization [7], [8]. Recently, DIC has also been extended to 3D volumetric image correlation [9], [10] by 3D imaging techniques such as X-Ray Computed Tomography (CT) or X-Ray Micro-Computed Tomography (MCT) [11].

In the present work, we present the results on the studying the initiation and evolution of strain localization in a rock analogue (granular, frictional, dilatant, and cohesive) material GRAM1 under triaxial axisymmetric compression at a confining pressure  $P_c$  corresponding to the transition from brittle faulting to ductile deformation  $P_{bdt}$ . This material has



the strength two orders of magnitude smaller than that of the most rocks, so it provides considerable technical/methodological advantages compared to the rock tests. In the context of the present work, this is particularly important as for  $P_c$  not exceeding a few mega-Pascals, the pressure cell can be made of transparent plastic allowing thus a direct observation of the sample surface during loading. From the images capturing the sample during loading, we apply the DIC technique to obtain the displacement and strain fields during the sample deformation.

## II. MATERIAL AND EXPERIMENTAL PROCEDURE

### A. GRAM1 Material characterization

The experiments presented in this article were carried out on GRAM1 (Granular Rock Analogue Material 1) which are fabricated from a finely ground powder of  $\text{TiO}_2$  with the average grain size of  $\sim 0.3 \mu\text{m}$ . The powder is subjected to the hydrostatic pressure of  $P_{\text{fab}} = 2 \text{ MPa}$  at which the grains are bonded one to another. An extensive program of GRAM1 stress-strain measurements in axisymmetric compression and extension tests at different  $P_c$  below  $P_{\text{fab}}$  shows that this

material has frictional, cohesive, dilatant properties and mechanical behavior very similar to those of hard rocks (Fig.1) [12], [13].

GRAM1 produces the same macro failure features as rocks, from very brittle splitting to shear and compaction bands with  $P_c$  increase, but is about 2 orders of magnitude less strong and less rigid than rocks. Both the friction coefficient  $\alpha$  and the dilatancy factor  $\beta$  reduce with  $P_c$  from positive to negative. GRAM1 presents a brittle-ductile transition behavior at  $P_c = P_{\text{bdt}} \approx 0.3 \text{ MPa}$ . The main physical and mechanical properties of GRAM1 are presented in Table 1.

TABLE I. MAIN PROPERTIES OF GRAM1 [18]

Properties	GRAM1
Porosity (n)	0.57
Density	1723 $\text{kg/m}^3$
Young's modulus $E$	$6.7 \times 10^8 \text{ Pa}$
Poisson's ratio $\nu$	0.24

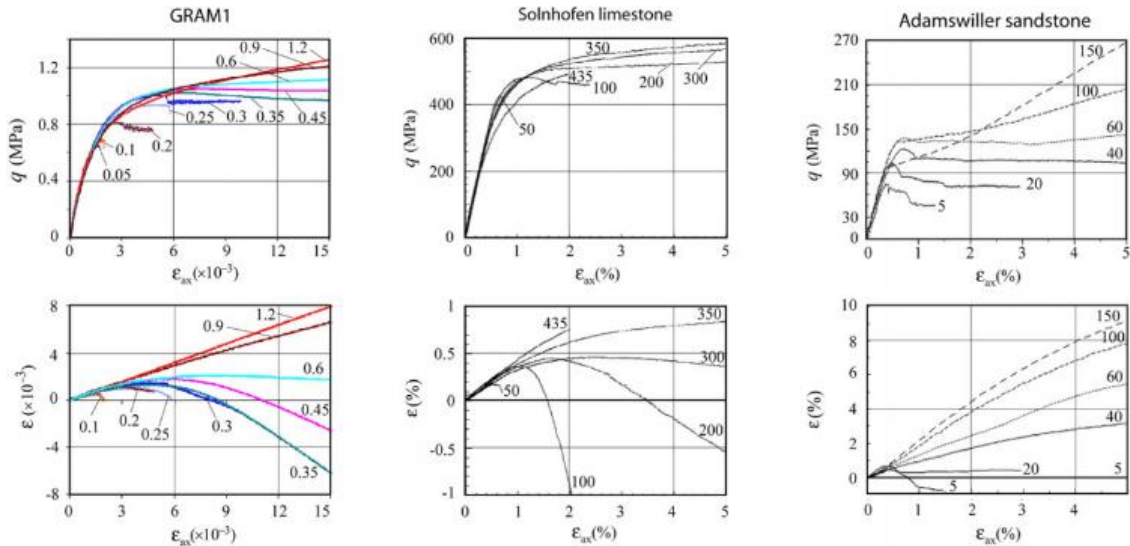


Fig. 1. Comparison of  $q(\epsilon_{ax})$  and  $\epsilon(\epsilon_{ax})$  curves for GRAM1 and rocks for compression tests. The plots for the rocks are from: Solnhofen limestone, Adamswiller sandstone [12].

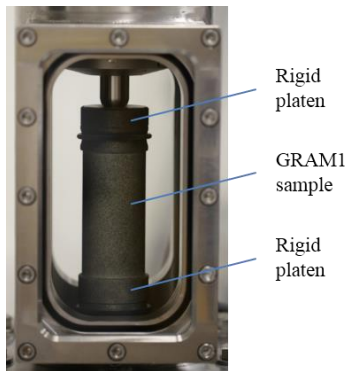


Fig. 2. Cylindrical GRAM1 sample jacketed with a very thin transparent latex film in axisymmetric compression tests

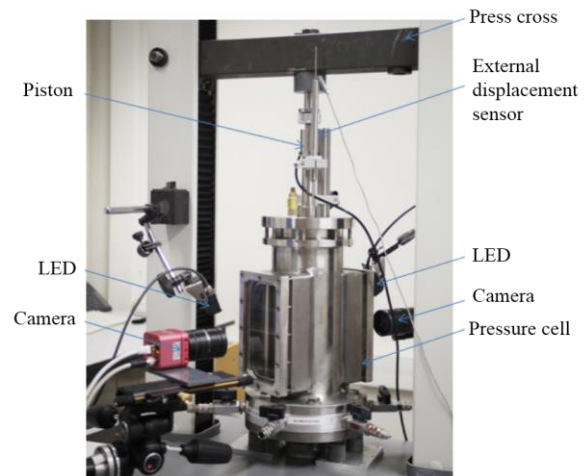


Fig. 3. Photo of the experimental setup

## B. Sample and Experimental setup

### 1) Sample

The testing was carried out on cylindrical test samples of about 80 mm in height and 40 mm in diameter. For DIC analysis, a speckle-like pattern is sprayed on a very thin ( $\sim 300 \mu\text{m}$ ) jacket to provide the surface with characteristic features. The droplets should be of an almost uniform size (of a few pixels, in our case about of 3-10 pixels), and sufficient density (Fig.2). The latter helps to increase the accuracy of DIC measurements [14].

### 2) Experimental setup

Fig.3 shows the experimental setup using the new pressure cell developed in Géoazur laboratory [15]. This new pressure cell with two transparent plane rectangular windows allow direct observations of the sample surfaces during loading. The sample and two rigid platens at the ends are jacketed and placed into a pressure cell of 2 MPa capacity filled with clean water and subjected to the confining pressure  $P_c$  of 0.3 MPa by a 3 MPa capacity-pressure generator (the accuracy is of  $4.5 \times 10^3 \text{ Pa}$ ). The sample is then loaded in the axial direction by a piston with a velocity of  $10^{-6} \text{ m.s}^{-1}$ , the axial stress induced is measured by an internal force sensor of 8 kN capacity with 0.1% precision. The deviatoric stress is computed as the major principal stress minus the minor principal stress  $q = \sigma_1 - \sigma_3$ .

The sample deformation was recorded throughout each test using two digital cameras with high resolution of  $2752 \times 2204$  pixels (with a pixel size of about  $32 \times 32 \mu\text{m}^2$ ) set up on two sides of plane windows of the pressure cell. The lighting system consists of four LED bulbs; two bulbs for each side are fixed on the top and bottom of window with a  $45^\circ$  inclination. To minimize spatial and temporal differences in the DIC uncertainties, it is important to align the axes of the cameras perpendicularly to these plane windows and keep a diffuse, homogeneous and unchanging illumination of the sample during the exposure time. During loading, images covering on the whole sample height are taken simultaneously at a frequency of 10 Hz; this allows investigating deformation bands from the initiation to their evolutions occurred in 0.1 s.

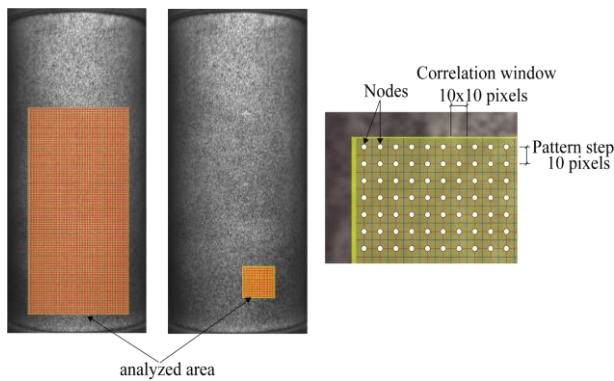


Fig. 4. Sample and the analyzed areas defined for computing the strains at different scales; Correlation parameters used in this study

### 3) Digital Image Correlation technique

Digital Image Correlation (DIC) is a method for non-contact full field kinematics measurement of planar or non-planar surfaces undergoing deformation. This method is based on the comparison of a reference image with an image at a deformed stage. Numerous DIC softwares have been developed, either academically (7D, Correla, CorrelManuV, KelKins, COSI-Corr...) or commercially (Aramis 2D, VIC-2D...). In this work, we use a software package 7D, which was developed at the Université de Savoie [16], for sub-pixel correlation.

A square grid is firstly defined in the analyzed area on a reference image of a sample and one or a set of sample images at later deformation stages. Around each grid node, a square correlation window (or subset) is then defined. In our analysis, the grid element size and correlation window size are both  $10 \times 10$  pixels (Fig.4). The DIC algorithm is designed to find the most similar subset between the reference and deformed images by optimizing a correlation coefficient  $C$  defined as (1):

$$C = 1 - \frac{\sum_{i \in D} (f(X_i) - \bar{f}) \times (g(x_i) - \bar{g})}{\sqrt{\sum_{i \in D} (f(X_i) - \bar{f})^2 \times \sum_{i \in D} (g(x_i) - \bar{g})^2}} \quad (1)$$

Where  $D$  is the subset over which  $C$  is computed.  $X_i$  and  $x_i = \phi_0(X_i)$  are respectively the coordinates (in pixels) of homologous points in the reference and deformed images.  $\phi_0$  is the transformation function.  $f(X_i)$  and  $g(x_i)$  are respectively the grey levels of the point  $i$  in the reference image and in deformed images.  $\bar{f}$  and  $\bar{g}$  are respectively the averages of the grey levels of all the pixels in subset of the reference and deformed images.  $C$  varies from 0 for perfect correlation, to the value 1 for no match at all.

The difference in the positions of grid nodes (centers of correlation windows) between the reference and deformed images yields a displacement vector, from which strain tensors might be calculated. Strains are calculated, in this case, using large strain Green Lagrange formulation.

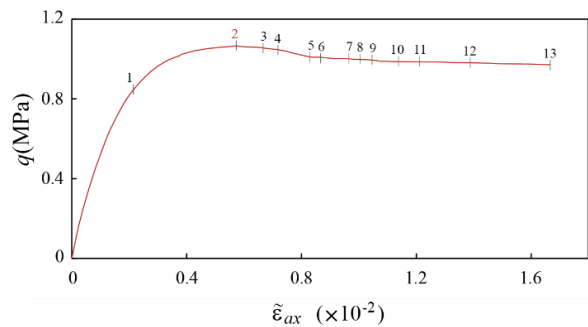


Fig. 5. Stress-strain curves  $q(\bar{\epsilon}_{ax})$  from the test C0.3-12T.  $\bar{\epsilon}_{ax}$  is the average of the axial DIC strains over the total height of the sample

### III. RESULTS AND DISCUSSION

In this section, the analysis of two representative tests conducted under the same condition, but different total shortening 1.9% and 8.1% respectively, will be presented.

#### A. Band initiation and propagation

Fig.5 presents the stress-strain curve for the presented test C0.3-12T which have clear peak succeeded by a smooth stress reduction and then, by a stress plateau. The nominal axial strain  $\tilde{\epsilon}_{ax}$  calculated from the DIC strains corresponds very well to the internal LVDT measurements [15] where the LVDT sensors were glued directly to the jacket of sample [12]. From 17000 photographs, 13 photographs corresponding to key moments in the test were selected for the DIC analysis. In order to study the evolution of the band network, we analyze  $\epsilon_{ax}$ , maximum shear ( $\gamma_m = \epsilon_{ax} - \epsilon_c/2$ ), and the incremental shear  $\Delta\gamma_m$  strains ( $\epsilon_{ax}$  and  $\epsilon_c$  are the axial and circumferential/horizontal strain).

From the evolution of accumulated shear strains  $\gamma_m$  in the test C0.3-12T presented in the Fig.6 for 13 representative points indicated on the stress-strain curve, the onset of strain localization can be seen to appear on the lower right corner of

the left hand side at point/stage 3. The initiated deformation band I form early, near the stress peak ( $q = 1.06$  Mpa), and then propagates rapidly with the average rate  $V_b$  of  $5 \times 10^{-1}$  mm/s (the rate of the sample shortening  $V_c$  is  $10^{-3}$  mm/s). Two other deformation bands (II) and (III) are initiated at stages 4 and 5 and then propagate with the rate of about  $3 \times 10^{-1}$  mm/s, which is much lower than the velocity of the first band propagation. These master bands are the most developed and remain active throughout the whole deformation history. Along with the master bands, secondary bands (bands a, b, c, d, g, e, f, h, i) form respectively during the deformation history (for more detail in [15]). The orientation of the shear bands with respect to the loading axis remains stable over the whole loading stage. This angle is of about  $34 \pm 1^\circ$ .

Fig.7 shows the evolution of accumulated axial strains  $\epsilon_{ax}$  in the case of large shortening (8.1%) C0.3-8T. Network of conjugate bands is formed by two families of deformation bands. Each family consists of both master bands and secondary bands. The master bands form quite early (from point/stage 3) and remain active during the deformation process. Along with these master bands, secondary bands form sequentially and network of conjugate bands therefore becomes denser.

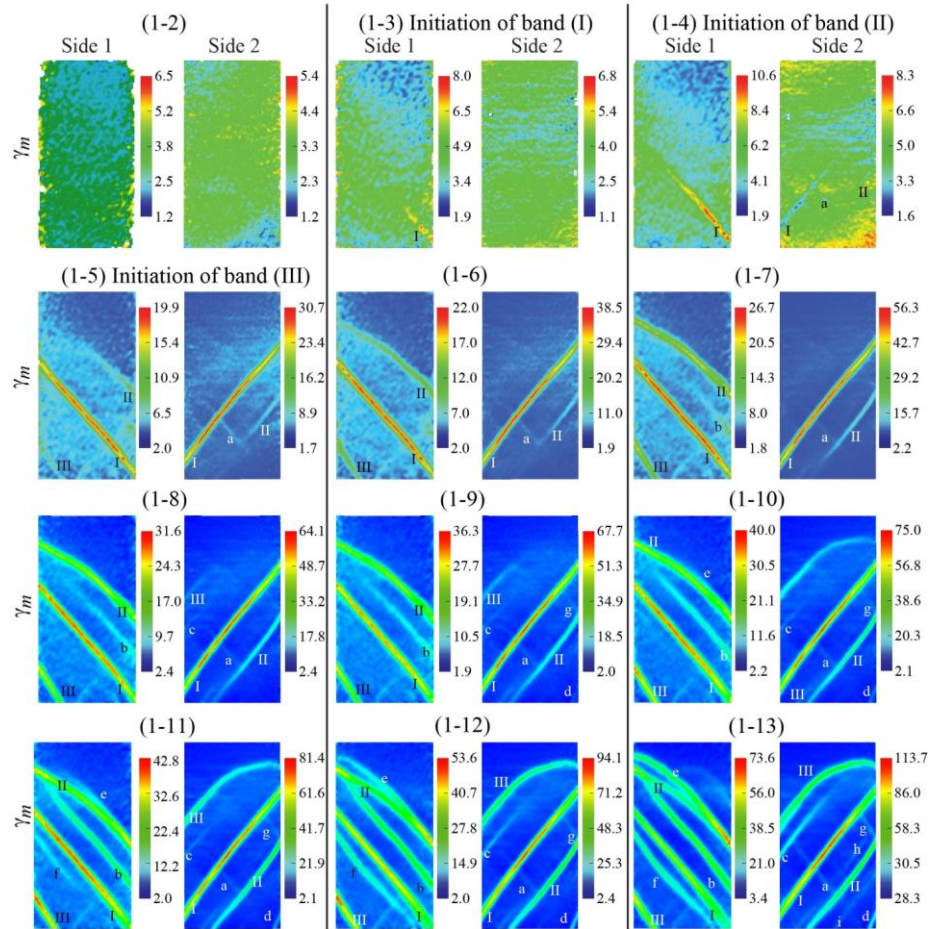


Fig. 6. Shear strains fields  $\gamma_m$  from DIC analyses. For each point, two images corresponding to the two opposite sides present strains  $\gamma_m$  accumulated between point 1 and points n, where  $n=2, 3 \dots 13$ . The numbers on color palettes are the strain values multiplied by  $10^3$  [15]



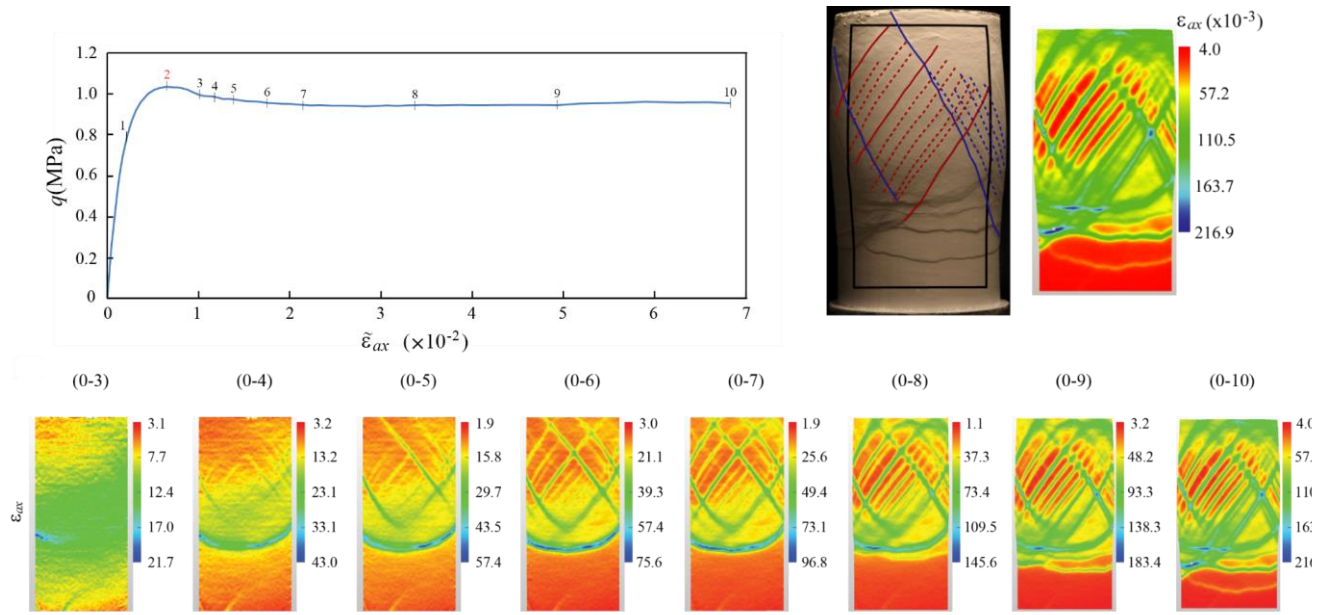


Fig. 7. Evolution of the  $\varepsilon_{ax}$  obtained from DIC in the test C0.3-8T. The numbers on the stress-strain curve from this test (in the upper left part corner of the figure) indicate the points for which the strains are shown. The number on color palettes are the strain values multiplied by  $10^3$

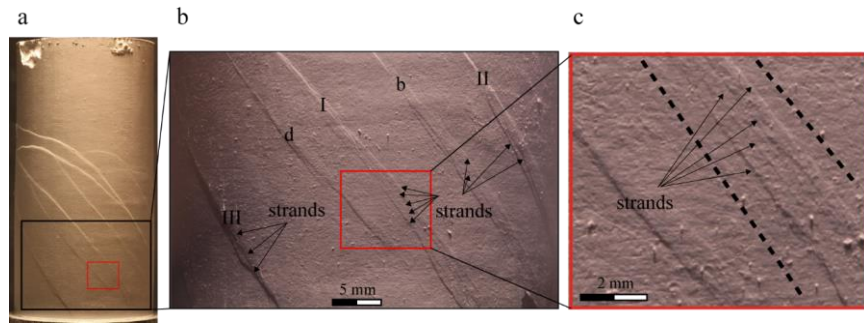


Fig. 8. Photos of post-mortem sample surface from test C0.3-12T zoomed at different scales: (a) the whole sample from the side 1, (b) in the black rectangle, (c) in the red rectangle.

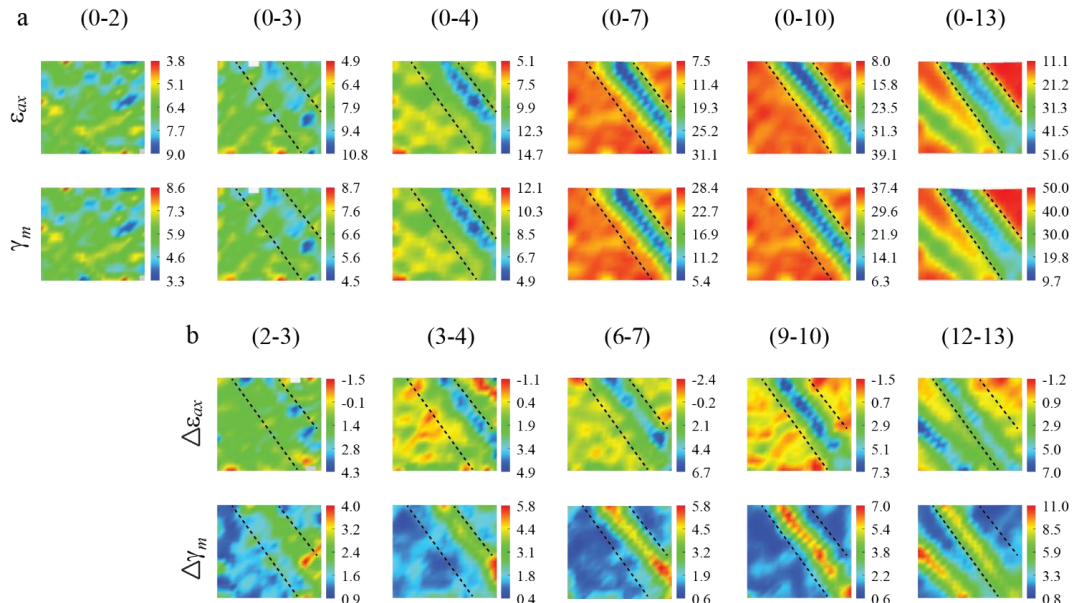


Fig. 9. Evolution of (a) the  $\varepsilon_{ax}$ ,  $\gamma_m$  fields and (b) incremental axial strains  $\Delta\varepsilon_{ax}$ , incremental shear strains  $\Delta\gamma_m$  obtained from DIC in the test C0.3-12T for the analyzed area defined in Fig.4 and Fig.8c. The number on color palettes are the strain values multiplied by  $10^3$

### B. Contribution of set of sub-parallel strands to band's thickness

The Fig.8 clearly shows that the evolution of an individual band includes its along-strike propagation and thickening. The thickness of the band increases progressively during the test and is contributed by a series of small bands, known as *strands*. Photos of the surface of a master band show the contribution of strands to final thickness of the master band (limited between two dotted lines) (Fig.8). These strands are quasi-sub parallel and formed sequentially, which is well seen on the incremental strain maps in Fig.9.

Fig.9a presents the evolution of the band (I)'s formation through the cumulative strain fields at six different levels of loading in the small area marked in red rectangle in Fig.8c. These cumulative strain fields are computed in comparison to the reference image at the initial stage (0). Two dotted lines present the final thickness of the band (I) at the end of the test, they therefore reveal the increasing of the band's thickening over time. Thus, by comparison of evolution of incremental strain fields in Fig.9b, it shows the process of forming of subsequent strands in this shear band zone. At step 2-3, the first strand initiates on the upper zone near the dotted line on the right side, and then the other strands in turn appear and

shift to the bottom dotted line (steps 3-4, 6-7, 9-10, 12-13). This allows identifying the evolution of the strands in deformation band in the analyzed zone on post-mortem sample (Fig.8c).

The thickness of the individual band/stand measured from DIC data  $d_D$  is thicker than that measured by zooming on the high-resolution photos of the post-mortem sample  $d_R$ . The actual thickness  $d_R$  of the incipient band is around 0.1 mm while  $d_D$  is about two grid elements ( $\sim 0.64$  mm) since the size of DIC grid element is 10 pixels (Fig.10b). When  $d_D$  is bigger than the grid element size,  $d_D$  approaches  $d_R$ . Fig.10 shows that both  $d_R$  and  $d_D$  of master band are about 2.3 mm.

As shown in Fig.11, in the case of 8.1% axial shortening, we can observe a network of dense conjugate bands in which each shear band zone (master bands as well as secondary bands) is contributed by a set of sub-parallel strands to the shear direction. The contribution of these strands to the main shear bands makes their orientation lightly decrease to of about  $32 \pm 2^\circ$  with respect to the axial principal stress. The thickness of these shear bands  $d_R$  (master bands as well as secondary bands) is about 2-3 mm, corresponds well to one measured from DIC data  $d_D$ .

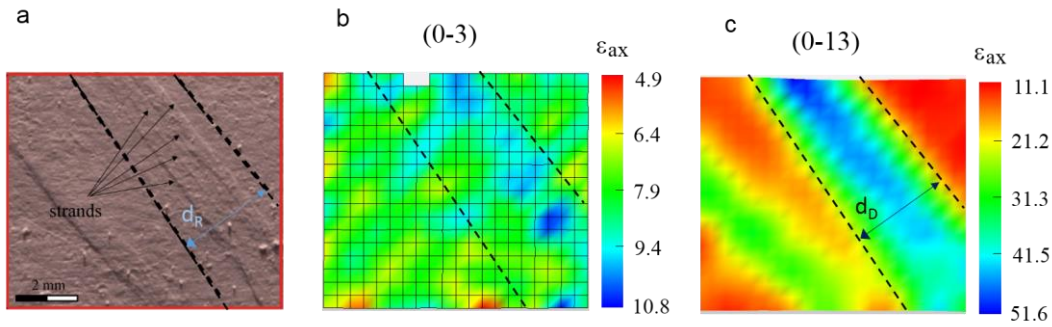


Fig. 10. (a) The thickness of master band  $d_R$  measured by zooming on the photo of post-mortem sample surface from test C0.3-12T, (b) (c) axial strain fields  $\epsilon_{ax}$  obtained from DIC at the deformation stages 3 and 13, respectively.  $d_D$  is the thickness of master band measured from DIC data.

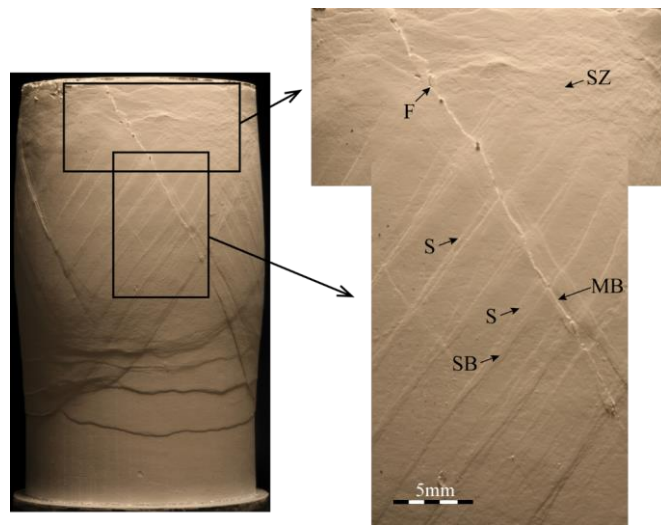


Fig. 11. Photo taken the whole sample from the test C0.3-8T and zoomed on the deformation band network. Different types of deformation band: major deformation band (MB), secondary deformation band (SB), strands contributing to the principle band (S), band saturation zone (SZ), open fissure (F)



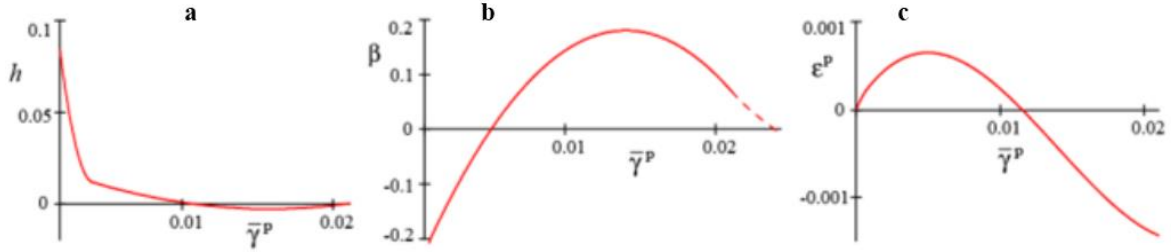


Fig. 12. Plots of constitutive functions for GRAM1 from [13]. (a)  $h(\bar{\gamma}^p)$ , (b)  $\beta(\bar{\gamma}^p)$ , (c)  $\varepsilon^p(\bar{\gamma}^p)$  at  $\sigma_m = \sigma_{\text{bdt}} = 0.653$  MPa respectively.

### C. Discussion

The results obtained from DIC analyses and post-mortem sample observation show the sequential formation of deformation bands. The evolution of each band includes its along-strike propagation and the forming of sets of sub-parallel strands to the shear direction contributing the thickness of the main shear band.

One explanation for deformation localization and the formation of band network at brittle-ductile transition is based on the theory of strain hardening materials, where the material failure is defined not only through the stress-strain evolution, but also through inelastic straining. That means considering the evolution the normalized hardening modulus  $h$  and dilatancy factor  $\beta$  with mean stress  $\sigma_m$  and equivalent inelastic shear strain  $\bar{\gamma}^p$ . From analysis of extended experimental data set for GRAM1 and data for Tavel limestone and Solnhofen limestone, Mas and Chemenda [13] showed for the first time the yield surfaces in the space  $(\bar{\tau}, \sigma_m, \bar{\gamma}^p)$ , where  $\bar{\tau}$  is the von Mises stress and  $\bar{\gamma}^p$  is the equivalent inelastic shear strain obtained by separation of elastic and inelastic strain. Their analysis confirmed that the deformation localization and rupture usually occurs when the normalized hardening modulus  $h(\sigma_m, \bar{\gamma}^p) = [\partial \bar{\tau}(\sigma_m, \bar{\gamma}^p) / \partial \bar{\gamma}^p] / G$  is close to zero ( $G$  is the shear modulus). The inelastic deformation starts well before the peak stresses and is characterized by positive hardening  $h > 0$ . Softening  $h < 0$  and dilatancy  $\beta < 0$  are correlated after the peak stresses, then correlation between hardening and compacting is observed at high mean stress  $\sigma_m$ ,  $\beta$  and  $h$  become positive (hardening).

At brittle-ductile transition, the mean stress at deformation localisation is located between the hardening and softening deformation regimes,  $\sigma_m = 0.653$  MPa in the reference test [15]. At this  $\sigma_m$ ,  $h$  changes from positive to negative value and then again increases close to 0. Therefore, one can assume that the first band initiates  $h < 0$  and evolved along the same band until  $h$  became positive.  $\beta$  respectively first increases from negative to positive values with  $\bar{\gamma}^p$  and then reduces to negative values again. As mentioned, the compaction is associated with the hardening, whereas the dilatancy is associated with the softening. The formation of band network occur thus at the transition from positive (softening) to negative (hardening)  $\beta$  with  $\bar{\gamma}^p$  growth. In consideration of  $\varepsilon^p(\bar{\gamma}^p)$  function, the formation of the first band (band I) in the reference test occurs between points 2 and 3 (on the stress – strain) where  $\bar{\gamma}^p \approx 0.016$ . The corresponding  $\beta$  is positive (close to the maximum value) and  $h$  is negative. After this point, the  $\bar{\gamma}^p$  increases rapidly, but then slows down again when  $\beta$  becomes negative and  $h$  is positive at  $\bar{\gamma}^p = 0.02$  (Fig.12). The material within the band therefore becomes stronger with deformation, it thus makes continued localization at the same location difficult. In contrast, stress is

locally enhanced in the surrounding regions which have been mechanically weakened and at a certain point it leads to form a new band in this weakened region.

### IV. CONCLUSION

The experimental methodology, combining both DIC analyses, post-mortem sample observation, allows us to detect the initiation of localisation deformation bands and their evolution during the loading and to understand their sequential activity at the macroscopic sample scale. The deformation bands do not form simultaneously but sequentially while they are not able to be identified by only post-mortem sample observation. Furthermore, DIC analysis allows giving some constraints on the velocity propagation of the deformation bands and showing the forming of sets of sub-parallel strands which contributes to the final thickness of the band. Both the dilatancy factor  $\beta$  and hardening modulus  $h$  evolve with  $\bar{\gamma}^p$ . The strain localisation occurs in the dilatant ( $\beta > 0$ ) deformation regime. During the band evolution, the deformation within it becomes compactive ( $\beta < 0$ ), which leads to widen the band and the formation of new bands. These experimental findings are close to that documented in the natural (geology) cataclastic bands [17], [18]. The process of strand network's formation contributing to the total band thickness may propose an explanation to the evolution of cluster network in field [19].

### ACKNOWLEDGMENT

This work was supported by the Côte d'Azur Observatory, the Region Provence Alpes Côte d'Azur and GeoFrac-Net Consortium.

### REFERENCES

- [1] T. F. Wong, C. David, W. Zhu, "The transition from brittle faulting to cataclastic flow in porous sandstones: Mechanical deformation", *Journal of Geophysical Research: Solid Earth*, 1997, 102.B2, pp. 3009–3025.
- [2] J. D. Byerlee, "Brittle-ductile transition in rocks", *Journal of Geophysical Research*, 73.14, 1968, pp. 4741–4750.
- [3] P. Bésuelle, J. Desrues, S. Raynaud, "Experimental characterisation of the localisation phenomenon inside a Vosges sandstone in a triaxial cell", *International Journal of Rock Mechanics and Mining Sciences*, 37.8, 2000, pp. 1223–1237.
- [4] J. W. Rudnicki, J. R. Rice, "Conditions for the localization of deformation in pressure-sensitive dilatant materials", *Journal of the Mechanics and Physics of Solids*, 23.6, 1975, pp. 371–394.
- [5] N. Lenoir, M. Bornert, J. Desrues, P. Bésuelle, and G. Viggiani, "Volumetric digital image correlation applied to x-ray microtomography images from triaxial compression tests on argillaceous rock", *Strain*, 2007, 43.3, pp. 193–205.
- [6] P. Bésuelle, G. Viggiani, N. Lenoir, J. Desrues, M. Bornert, "X-ray micro C for studying strain localization in clay rocks under triaxial compression", *Adv X-ray Tomogr Geomater*, 2010.
- [7] P. Doumalin, "Characterisation of the strain distribution in heterogeneous materials", *Mécanique Ind.*, 4(6):607–617, 2003.

- [8] P. Bésuelle, P. Lanatà, “A new true triaxial cell for field measurements on rock specimens and its use in the characterization of strain localization on a vosges sandstone during a plane strain compression test”, *Geotechnical Testing Journal*, 39(5), pp.879–890, 2016.
- [9] M. Bornert, J. M. Chaix, J. C. Dupré, T. Fournel, D. Jeulin, H. Moulinec, “Mesure tridimensionnelle de champs cinématiques par imagerie volumique pour l’analyse des matériaux et des structures, Instrumentation”, *Mes Métrologie*, 4(3-4):43-88, 2004.
- [10] N. Lenoir, M. Bornert, J. Desrues, P. Bésuelle, and G. Viggiani, “Volumetric digital image correlation applied to x-ray microtomography images from triaxial compression tests on argillaceous rock”, *Strain*, 43(3), pp.193–205, 2007
- [11] F. Hild, E. Maire, S. Roux, J. F. Witz, “Three-dimensional analysis of a compression test on stone wool”, *Acta Materialia*, 57(11), pp.3310–3320, 2009.
- [12] S. H. Nguyen, A. I. Chemenda, J. Ambre, “Influence of the loading conditions on the mechanical response of granular materials as constrained from experimental tests on synthetic rock analoguematerial”, *International Journal of Rock Mechanics and Mining Sciences*, 48(1), pp.103–115, 2011.
- [13] D. Mas, A. I. Chemenda, “An experimentally constrained constitutive model for geomaterials with simple friction–dilatancy relation in brittle to ductile domains”, *International Journal of Rock Mechanics and Mining Sciences*, 77, pp.257–264, 2015.
- [14] J. Dautriat, M. Bornert, N. Gland, A. Dimanov, J. Raphanel, “Localized deformation induced by heterogeneities in porous carbonate analysed by multi-scale digital image correlation”, *Tectonophysics*, 503(1–2), pp.100–116, 2011.
- [15] T. P. H. Tran, S. Bouissou, A. Chemenda, J. Ambre, P. Vacher, and P. Michel, “Initiation and Evolution of a Network of Deformation Bands in a Rock Analogue Material at Brittle-Ductile Transition”, *Rock Mechanics and Rock Engineering*, 52, pp.737–752, 2019.
- [16] P. Vacher, S. Dumoulin, F. Morestin, S. Mguil-Touchal, “Bidimensional strain measurement using digital images. Proceedings of the Institution of Mechanical Engineers”, Part C: *Journal of Mechanical Engineering Science*, 213(8), pp.811–817, 1999.
- [17] A. Aydin, A. M. Johnson, “Analysis of faulting in porous sandstones”, *J Struct Geol* 5:19–31, 1983.
- [18] E. Sallet, C. A. J. Wibberley, “Evolution of cataclastic faulting in high-porosity sandstone, Bassin du Sud-Est”, *J Struct Geol* 32(11):1590–1608, 2010.
- [19] S. Philit, R. Soliva, R. Castilla, G. Ballas and A. Taillefer, “Cluster of cataclastic deformation bands in porous sandstones”, *Journal of Structure Geology*, 114, pp.235:250, 2018.

# Effects of the Relative Humidity on the Performance of Thermoelectric Freshwater Generator using Solar Power Source

Le Minh Nhut  
Department of Thermal Engineering  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
nhutlm@hcmute.edu.vn

Dang Thi Truc Linh  
Department of Thermal Engineering  
Cao Thang Technical College  
Ho Chi Minh City, Vietnam  
dangtruclinh77@gmail.com

**Abstract**— The main objective of this study is to evaluate the effects of the relative humidity (RH) on the performance of thermoelectric freshwater generator using solar power source. The system was designed, constructed and experimented under real weather conditions in Southern Vietnam. The system mainly consists of four thermoelectric coolers, the four axial fans, the four hot side heat sink and cooling fans, and four cold side heat sinks, two solar panels, charge controller and condensate water box. The experimental results showed that the amount of collected water of the system was affected by changes in the RH, electric current, and air mass flow rate. The collected water reached its highest value of 215 ml while the value of the coefficient of performance (COP) was 0.73 at the RH of 80 %.

**Keywords**—thermoelectric cooler, relative humidity, heat sink, solar cell

## I. INTRODUCTION

During the last few decades, the growth of the population and the industry has led to a rapid increasing in pollution from CO<sub>2</sub> emissions and global climate change. This is also caused of the increasing in the prolonged drought and severe water scarcity in many areas on the world. Mekonnen [1] reported the facing of water scarcity of people on the world and proposed the solution to increase the efficiency of of water using, and better sharing of the limited freshwater resources will be key in reducing the threat posed by water scarcity on biodiversity and human welfare. Vian et al. [2] presented the effect of heat transfer and phase change during condensation and evaporation of low power (100W) thermoelectric dehumidifier with two stages and. The results indicated that the obtained value of COP is 0.8 at atmospheric conditions of 32 °C and 90% humidity. Udomsakdigool et al. [3] developed the new hot heat sink with a rectangular fin array for thermoelectric dehumidifiers and they concluded that the efficiency was 95% and COP of TE dehumidifier was 0.88 at electric current of 1.9 A and voltage of 12 V. Tan [4] made an air-based water collector that uses thermoelectric cooling device. They concluded that at the value of the humidity of 82%, the amount of condensed water was 16.8 ml/h. Joshi et al. [5] proposed the thermoelectric fresh water generator with 0.7 m long cooling channel and ten thermoelectric modules. It was reported that by using of internal heat sink, the quantity of water generated per 10 hours increases by 81% as compared without internal heat sink. Yao et al. [6] have developed the thermoelectric dehumidifier with two cold side

heat sinks and two hot-side heat sinks. The experimental results indicated that the moisture removal rate of the novel prototype is up to 33.1 g/h and the corresponding COP is 0.75 when the air flow velocity of air duct is 1.74 m/s. The performance of thermoelectric exhaust heat recovery system considering different heat source's fin arrangements also reported by Sofyan and et.al [7]. Shourideh et al. [8] developed the atmospheric water generator using thermoelectric cooling effect with the cold side extended surface design optimization. They have also demonstrated the lowest specific energy consumption for both relative humidity levels (60% and 80%) comparison various the atmospheric water generator using thermoelectric cooling effect available in literature before. In this paper, the thermoelectric freshwater generator using solar power source based on the atmospheric water generator using Peltier effect model which was proposed by Shourideh et al. [8] has been extended and modified for experiments to evaluate the effect of the relative humidity on the performance of thermoelectric freshwater generator using solar power source under real weather conditions of South Vietnam.

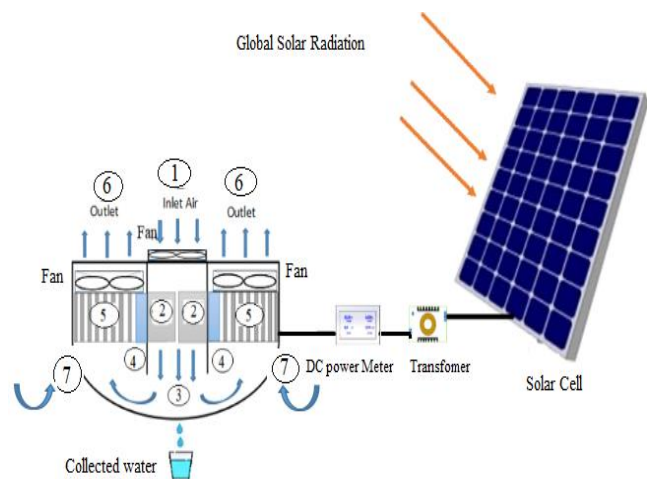


Fig. 1. The schematic diagram of thermoelectric freshwater generator using solar power source: (1), (3), (6) air temperature and humidity sensors, (2) cold side heat sink, (4) thermoelectric cooler, (5) hot side heat Sink, (7) air intake.

## II. SYSTEM DESCRIPTION

### A. Experimental setup

Thermoelectric freshwater generator using solar power source has been designed, constructed and experimented under real weather conditions in Southern Vietnam to evaluate the effects of the relative humidity (RH) on its performance. The schematic diagram of thermoelectric fresh water generator using solar power source is shown in Fig.1. The system consists of four thermoelectric coolers, the four axial fans and which is mounted at the entry of channel which helps in generating the suction of humidity air into the system, the four hot side heat sinks and cooling fans, the four cold side heat sinks, solar panel, storage battery, charge controller and condensate water box. The detail components of the system are shown as located in Fig. 2, and the specifications of the components are given in Table 1. The thermoelectric fresh water generator operates based on the thermoelectric cooling effect by condensing the vapor from the ambient humidity air. Thermoelectric freshwater generator has two sides (cold side heat sink and hot side heat sink), and when the DC current flows through the system, it brings heat from one side to other, so that one side gets cooler while the opposite one gets hotter.

TABLE I. TECHNICAL CHARACTERISTICS OF THE SYSTEM

Peltier module	Model: TEC1-12710
	Operating voltage: 12V
	Maximum voltage: 15V
	Maximum current: 10A
	Cooling capacity (W) : 90W
	Temperature difference: 66°C
Hot side heat sink	Model: Infinity Dark Fallen
	Heat dissipated capacity: 130W
	Voltage: 12V
	Current: 0.25A
	Heat sink material: aluminum fins and copper heat pipes
	Dimension: 120x120x25mm
	Spin speed: 1600RPM
	Max wind flow: 50CFM
Cold side heat sink	Dimension:40x45mm
	Fin thickness: 1mm
	Space between fins: 4mm
	Dimension: 60x60x15mm
axial fan	Max wind flow: 50CFM
	Voltage: 12V
	Current: 0.14A
	Electric power: 1.68W
Solar cell	Model:SHINSUNG E&G SS-BM300
	Peak power (Wp): 300
	Open circuit current: 39.10V
	Short circuit current: 10.04V
	Voltage at maximum power: 31.29V
	Current at maximum power: 9.59A

The operation of system can be described as follows. The humidity air enters the system from the top of center channel due to the suction generated by the axial fan. Then the humidity air starts goes through the thermoelectric cooling channel wherein heat is absorbed from the humidity air by peltier modules at the cold side heat sink, and then the humidity air is cooled and its temperature gradually reduces

to the dew point and then the condensation process of vapor stars. The water generation is collected at the bottom of the channel. The air then enters the hot side heat sink absorbed the amount heat rejected of the hot side and leaves to the outside environment.



Fig. 2. The photo of experimental setup

### B. Experimental procedure

In this work, the experiments were conducted in different work conditions under real weather conditions in South Vietnam with the change of ambient temperature, relative humidity, mass flow air and current. In these experiments, the ambient temperature ranges from 28 °C to 34 °C, and relative humidity was varied between 40% and 80%. The current of input power source were controlled at 6A, 8A, 10A, 12A respectively. The four different inlet air flow rates (20 kg/h, 40 kg/h, 60 kg/h, and 80 kg/h) are adjusted by the dimmer of axial fan. Several T-type thermocouples were used to measure temperatures of air in channel and outside environment (uncertainty of  $\pm 0.5\%$ ), and relative humidity was measured with a hygrograph (accuracy of  $\pm 3\%$ ). Anemometer of accuracy  $\pm 3\%$  is used to measure the mass flow rate of moist air flowing through the channel. The data was recorded every 5 min and saving in excel file for calculation.



### III. RESULTS AND DISCUSSION

The variation of water condensate versus time at different relative humidity is shown in Fig.3. As seen in Fig. 3, the amount of water condensate increased with an increase in the relative humidity. The total amount of water condensate in the 4h experiment was 37 ml, 68 ml, 97 ml, 137 ml and 187 ml when the relative humidity of the inlet air was equivalent to 40%, 50%, 60%, 70% and 80%, respectively. This can be briefly explained that when the air flow rate inlet the channel is constant, the increasing in relative humidity leads to a rapid reduction in the temperature of moist air at the cold side heat sink of the thermoelectric freshwater generator system.

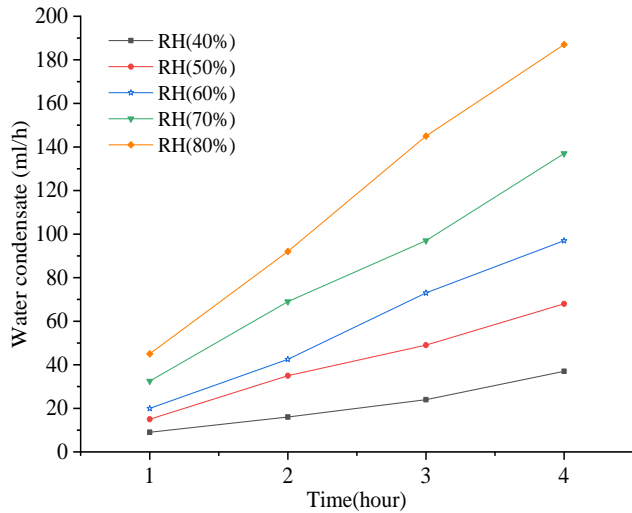


Fig. 3. Variation of water condensate versus time at different relative humidity

Fig. 4 shows the effect of relative humidity on COP of the thermoelectric freshwater generator system. As seen in Fig. 4, when the relative humidity ranges from 40% to 60%, the COP increases slowly from 0.14 to 0.37, and when the relative humidity ranges from 60% to 80%, the COP increases rapidly from 0.37 to 0.73. This is because the rise in relative humidity from 60% to 80% contributes to an enhancement in the heat transfer efficiency between the moisture air and the cold side heat sink. The maximum value of the COP is 0.73, which corresponds to the relative humidity of 80%.

The effect of mass flow rate on the water condensate at different relative humidity is shown in Fig. 5. As indicated in Fig. 5, of course, with the increase in the air mass flow rate, the generated water is also increased. As the same the air mass flow rate, the increase of the water condensate is proportion with respect to the increase of the relative humidity. This is due to the effect of the water vapour content in the moisture air. The maximum value of water condensate is 215 ml corresponds with the relative humidity and the air mass flow rate are 80% and 100 kg/h, respectively.

The effect of current on water condensate with different relative humidity is shown in Fig. 6. At the each level of the relative humidity, the amount of the water condensate increases in proportion to the current. This is because the

increase in the current leads to the increase in the heat absorbed by the cold side heat sink of the system.

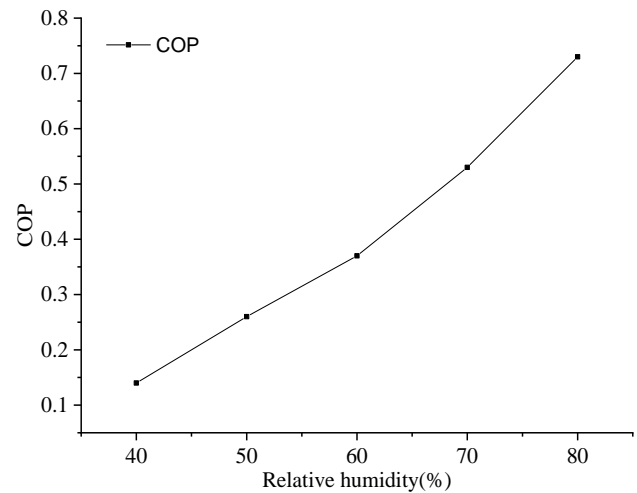


Fig. 4. Effect of relative humidity on COP of system.

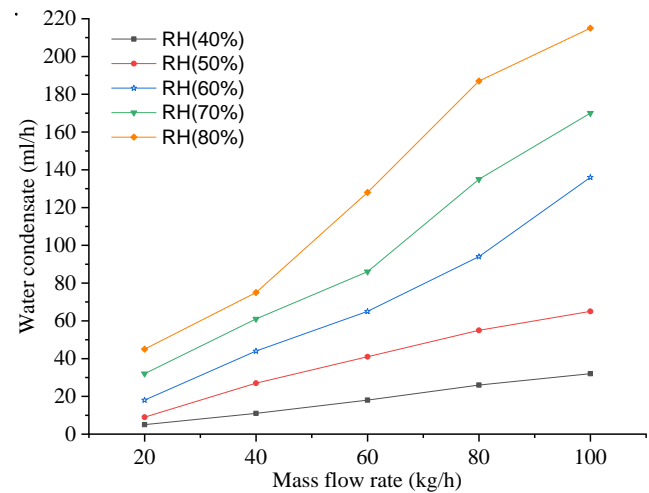


Fig. 5. Effect of air mass flow rate on water condensate with different relative humidity.

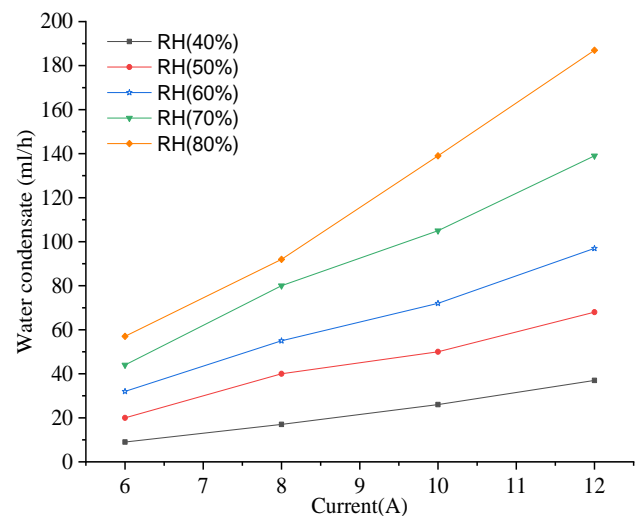


Fig. 6. Effect of current on water condensate with different relative humidity.



#### IV. CONCLUSION

An experimental investigation of the thermoelectric freshwater generator using solar power source was designed, fabricated, and tested under real weather conditions of South Vietnam with different RH, air mass flow rates, and current. The following conclusions can be drawn:

- (1) The amount of water condensate increased with the increase in the relative humidity and the maximum water generated is found to be 187ml corresponds RH is 80%
- (2) Increasing RH also leads to an increase of the COP system. The maximum value of the COP is 0.73, which corresponds to the relative humidity of 80%.
- (3) Increasing the air mass flow rates strongly affected the amount of water generated of system, and the maximum generated water was 215 ml with an air mass flow rate of 100 kg/h.
- (4) The amount of the condensate water increases in proportion to the electric current.

#### ACKNOWLEDGMENT

This work was supported by Ho Chi Minh City University of Technology and Education, Vietnam.

#### REFERENCES

- [1] M. M. Mekonnen and A. Y. Hoekstra, "Four billion people facing severe water scarcity," *Science Advances*. Vol. 2, pp. 2–6, 2016.
- [2] J. G. Vian, D. Astrain and M. Dominguez, "Numerical modelling and a design of a thermoelectric dehumidifier," *Applied Thermal Engineering* Vol.22, , pp.407–422, 2002.
- [3] C. Udomsakdigool, J. Hirunlabh, J. Khedari and B. Zeghmami, "Design optimization of a new hot heat sink with a rectangular fin array for thermoelectric dehumidifiers," *Heat Transfer Engineering* Vol. 28, pp. 645–655, 2007.
- [4] F.L. Tan and S.C. Fok, "Experimental testing and evaluation of parameters on the extraction of water from air using thermoelectric coolers," *Journal of Testing and Evaluation*, Vol. 41, pp. 96–103 , 2013.
- [5] V.P. Joshi, V.S. Joshi, H.A. Kothari, M.D. Mahajan, M.B. Chaudhari, and K.D. Sant, "Experimental investigations on a portable fresh water generator using a thermoelectric cooler," *Energy Procedia*, Vol. 109, pp. 161–166, 2017.
- [6] Y. Yao, Y. Sun, D. Sun, C. Sang, M. Sun, L. Shen, and H. Chen, "Optimization design and experimental study of thermoelectric dehumidifier," *Applied Thermal Engineering* Vol. 123, pp. 820–829, 2017.
- [7] S. E. Sofyan, Muhajir, Khairil, Jalaluddin, and S. Bahri, "The performance of thermoelectric exhaust heat recovery system considering different heat source's fin arrangements," *Earth and Environmental Science*, Vol 463, 2020.
- [8] A. H. Shourideh, W. Bou Ajram, J. Al Lami, S. Haggag, and A. Mansouri, "A Comprehensive Study of an Atmospheric Water Generator using Peltier Effect," *Thermal Science and Engineering Progress*, 2018.
- [9] W. He, P. Yu, Z. Hu, S. Lv, M. Qin, and C. Yu, "Experimental study and performance analysis of a portable atmospheric water generator," *Energies*, Vol 13(1), 2019.
- [10] A. C. Sulaiman, N. A. M. Amin, M. H. Basha, M. S. A. Majid, N. F. B. M. Nasir and I. zaman, "Cooling performance of thermoelectric cooling (TEC) and applications: A review," *MATEC web of Conference* 225, 03021, 2018.
- [11] W. Nandy, S. Saha, S. Ganguly and S. Chattopadhyay, "A project atmospheric water generator with the concept of Peltier effect," *Int. Journal of Advanced Computer Research*, Vol 4(2), 2014.
- [12] Y. Wang, Y. Shi and D. Liu, "Performance analysis and experimental study on thermoelectric cooling system coupling with heat pipe," 10<sup>th</sup> Int. Symposium on Heating, Ventilation and Air Conditioning, ISHVAC2017, 19-22 October 2017, Jinan, China.

# Effect of Alkaline Solution and Curing Conditions on the Strength of Alkali – Activated Slag Mortar

Tai Tran Thanh

Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
taitt@hcmute.edu.vn

Tu Nguyen Thanh

Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tunt@hcmute.edu.vn

Hyug-Moon Kwon

Civil Engineering Department  
Yeungnam University  
Gyeongsan, Korea  
hmkwon@yu.ac.kr

**Abstract**—The aim of this work was to study the compressive strength and microstructure of alkali – activated slag mortar (AAS mortar) under different curing methods. Granulated blast furnace slag (GBFS) was activated by the combination of sodium silicate liquid ( $\text{Na}_2\text{SiO}_3$ ) and sodium hydroxide ( $\text{NaOH}$ ) with a silicate modulus (a  $\text{SiO}_2/\text{Na}_2\text{O}$  weight ratio) of 1 and different  $\text{Na}_2\text{O}$  concentrations of 4, 6, 8 % by slag weight. AAS mortar specimens were cured in the ambient condition ( $20 \pm 5^\circ\text{C}$ ,  $60 \pm 5\%$  RH) (RH: Relative humidity), in a dry oven at  $80^\circ\text{C}$ , and in the saturated limewater at  $80^\circ\text{C}$  for 24 hours. After curing, the specimens were kept in the ambient condition until the time of testing. Scanning electron microscopy (SEM) was used to assess the mortar microstructure. The results revealed that the  $\text{Na}_2\text{O}$  concentration of alkaline solution and curing method significantly affected the compressive strength of AAS mortar. Curing in the saturated limewater at the high temperature was found to be effective in promoting the mortar strength at later ages.

**Keywords**—alkali-activated slag, mortar, concentration, curing condition, strength, microstructure

## I. INTRODUCTION

Portland cement (PC) has been known as the main binder using for producing the concrete material, which is the most widely used material in the construction field. The Portland cement has a considerable environmental impact when its annual emission of  $\text{CO}_2$  into the air is 1.5 billion tonnes. In addition, the tremendous amount of natural resources and energy is consumed during this binder production process [1]. As a result, it is very necessary to replace Portland cement by the other binder which is more friendly to the environment.

In recent decades, alkali – activated material (AAM) has received the strong attention from many researchers due to its good mechanical and durability performance [2][3]. For AAM synthesis, some industrial aluminosilicate by-products, such as granulated ground blast furnace slag (also called slag), fly ash, metakaolin or rice husk ash could be activated by the highly alkaline solution ( $\text{pH} = 12 - 14.5$ ) [4]. Slag, the by-product from cast iron manufacture [5], is one of the most commonly materials using as the main raw material for AAM synthesis. When comparing with Portland cement, alkali – activated slag (AAS) was found to exhibit some advantages such as the rapid development of mechanical strengths [6][7], better durability in the aggressive chemical environment [8][9] or elevated temperature exposure [10][11], lower hydration heat. The major hydration product of alkali – activation slag is hydrated calcium silicate (C-S-H) with a poor crystalline form [12][13]. AAS binder has been expected

to be a potential material for Portland cement replacement [14].

Many studies concluded that the properties of AAS are considerably dependent on the alkaline solution [15][16][17] and curing condition [18][19]. It has been reported that the ingredient and concentration of the alkaline solution have the significant influence on the hydration kinetic of slag. When comparing with the sodium hydroxide activated slag, AAS material using sodium silicate as the activator possesses the much more slowly hydration process, but attaining the greater strength promotion at later – age [6][20][21]. In spite of the early strength gain, increasing the sodium hydroxide activator concentration often does not result in the AAS mechanical strength [14][20][22]. Additionally, the mechanical strength of AAS can be enhanced by raising the dosage of sodium oxide and silicate modulus ( $M_s$ , the mass ratio of  $\text{SiO}_2$  and  $\text{Na}_2\text{O}$ ) in the alkaline solution [15][17]. Several investigations have been carried out on the behavior of AAS cured in many different conditions. The AAS concrete or mortar can attain the high mechanical strength at the early age with curing at elevated temperatures [23]. The AAS microstructure formation can be also accelerated in this curing condition. Moreover, the curing regime with both high temperature and relative humidity was pointed out to lead to the superior performance of AAS concrete, followed by the ambient curing and saturated limewater curing [17]. So far, very little attention has been paid to the role of the other curing method using the high alkaline solution on the AAS behavior.

The purpose of this study is to explore the compressive strength and microstructure of AAS mortar using the saturated lime water with high temperature for curing. In addition, the influence of different concentration of  $\text{Na}_2\text{O}$  on the AAS mortar properties is also accessed in this investigation.

## II. MATERIALS AND METHOD

### A. Material characterization

Slag from Korea supply was used as the precursor for the alkali – activation reaction. The chemical composition of the raw binder was identified by using an X-ray fluorescence (XRF) method, and presented in Table I. In addition, the slag has a specific gravity of  $2.9 \text{ g/cm}^3$  and a Blaine surface area of  $435 \text{ m}^2/\text{kg}$ . For AAS synthesis, the mixture of sodium silicate liquid (water glass) and sodium hydroxide ( $\text{NaOH}$ ) was used as the activator. The sodium silicate liquid contains 26.4 %  $\text{SiO}_2$ , 8.2 %  $\text{Na}_2\text{O}$ , and 65.4 %  $\text{H}_2\text{O}$  by mass. The alkaline activator was made by dissolving the 97 – 98 % pure  $\text{NaOH}$  pellets in the sodium silicate liquid, and prepared prior to

mixing the mortar mixture 24 hours. The fine aggregate with the size smaller than 5 mm was derived from the natural sand, and has the fineness modulus of 2.45.

TABLE I. CHEMICAL COMPOSITION OF SLAG

Oxide composition	Slag (%)
SiO <sub>2</sub>	33.81
CaO	41.24
Al <sub>2</sub> O <sub>3</sub>	15.19
Fe <sub>2</sub> O <sub>3</sub>	0.41
MgO	5.54
SO <sub>3</sub>	2.51
Na <sub>2</sub> O	0.25
K <sub>2</sub> O	0.61
LOI	0.18

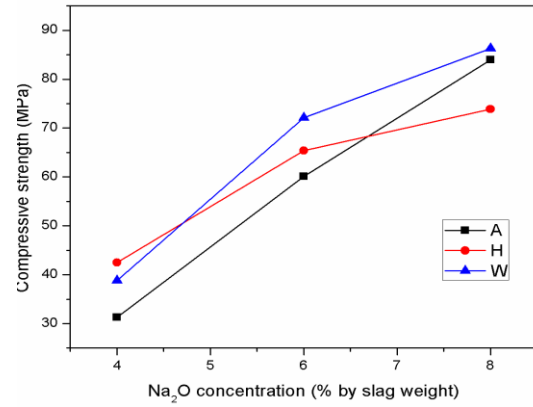
### B. Mixture proportion

The mortar mixture has the fine aggregate and slag mass ratio of 2.75. Mass portion of the alkaline activator portion was calculated based on the Na<sub>2</sub>O concentration (percentage of slag weight, and the silicate modulus of 1 (M<sub>s</sub>, a mass ratio of SiO<sub>2</sub> and Na<sub>2</sub>O). Three different mortar mixtures with the Na<sub>2</sub>O concentration of 4, 6, and 8 % by slag weight were named as A4, A6, and A8, respectively.

### C. Experimental program

For making the fresh AAS mortar mixture, ingredients were mixed by using a 5 L capacity planetary mixer. Then, this mixture was casted into the 50 mm cube triplicate molds in accordance with the compressive strength test, and kept in the ambient condition (20 ± 5 °C, 60 ± 5 % RH) for 24 hours. In the following step, the specimens were removed from their molds and stored in three various curing regimes for next 24 hours: in the laboratory environment (the ambient condition), in a dry environment with a temperature of 80 °C (in an oven), and in the saturated lime water (80 °C), named as A, H and W, respectively. After curing, the mortar samples were continuously kept in the ambient condition until time of testing. At 3, 7, 28, 56 and 120 curing days, the mortar compressive strength test was conducted by using the hydraulic machine according to standard ASTM C109 [24].

The mortar microstructure identification was determined on the chosen mortar fragments by using the scanning electron microscopy (SEM). After compression test, some selected fragments from the broken specimen were immersed in acetone for 3 days aiming to stop the further hydration, and then stored in a vacuum desiccator for next 1 day.

Fig. 1. Relation between Na<sub>2</sub>O concentration and AAS mortar strength.

## III. RESULTS AND DISCUSSION

### A. Influence of the alkaline solution

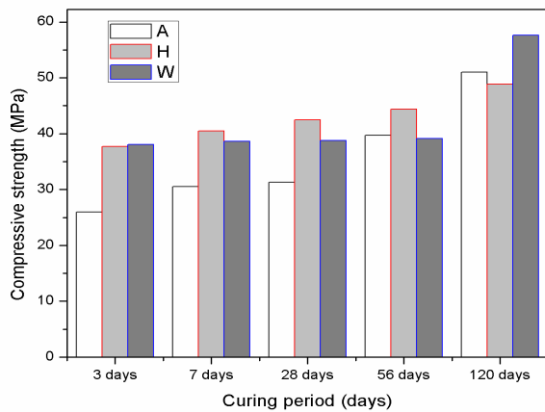
The compressive strength of AAS mortar using an activator with three Na<sub>2</sub>O dosage of 4 %, 6 % and 8 % by slag weight is showed in Figure 1. As shown in Figure 1, there was a significant correlation between the Na<sub>2</sub>O concentration of activator and the compressive strength of mortar. Increasing the concentration of Na<sub>2</sub>O can result in the mechanical strength gain. It can be explained by the acceleration of chemical reaction, caused by the rise of alkaline level in the activator [25][26]. Figure 1 reveals that there was a sharp rise in the compressive strength when the Na<sub>2</sub>O concentration increased from 4 to 6 %. As the dosage of Na<sub>2</sub>O was higher than 6 %, the strength gain was seen to reduce in specimens cured in the high temperature conditions (H and W method). For the ambient condition cured mortar, the reduction in strength gain at 28 days was found to be negligible when the dosage of Na<sub>2</sub>O increased from 6 to 8 %. Due to high sodium content, AAS matrix also performed the shrinkage deformation [26]. Consequently, the aforementioned strength gain was reduced when increasing the Na<sub>2</sub>O content from 6 to 8 %.

### B. Influence of the curing condition

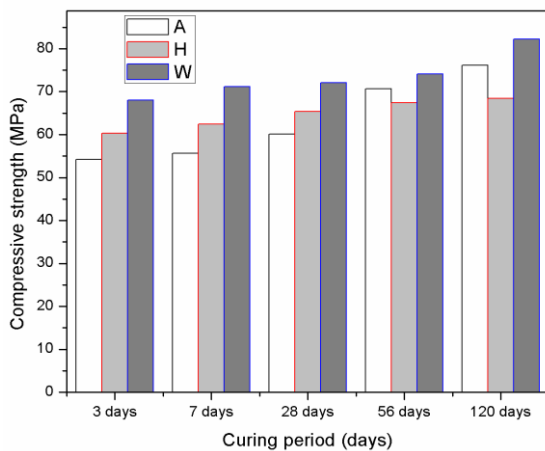
Figure 2 presents the experimental data on the compressive strength of AAS mortar cured with different three regimes. The strength of different three mortar mixtures was measured at 3, 7, 28, 56, and 120 curing days. From these graphs, it is noted that the mortar strength at the early age was improved by using the heat treatment (H and W method). For instance, at the age of 3 days, the H method resulted in the strength gain of 45, 11.3, and 1 % for specimen with the Na<sub>2</sub>O dosage of 4, 6, and 8 %, respectively. Interestingly, this strength gain of mortar was observed to increase to 46.5, 25.5, and 21.8 % when using the W curing method. Additionally, the AAS mortar exhibited the lower strength increase due to heat treatment with higher concentration of Na<sub>2</sub>O from 4 to 8 %.

It was noticeable from Figure 2 that there was a continual growth of compressive strength of all mixtures during the curing period. In the range from 3 days to 120 days, the A curing method led to the highest strength gain in AAS mortar specimens, followed by the W curing and H curing method. In spite of the low strength at early age, at the curing age of 120 days, AAS mortar cured in the ambient condition can attain the higher strength than specimen cured in an oven at 80 °C. The curing regime in saturated lime water at 80 °C was found

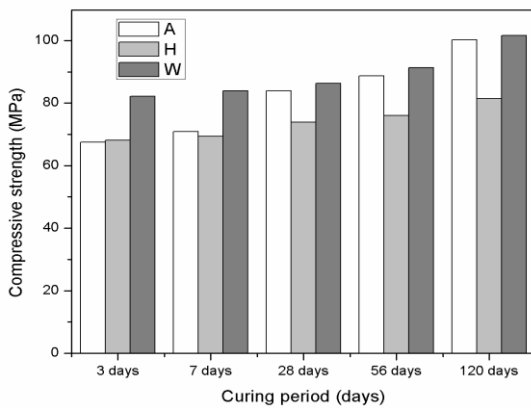
to be the most effective method when it can give the highest compressive strength for AAS mortar at 120 days.



(a)



(b)



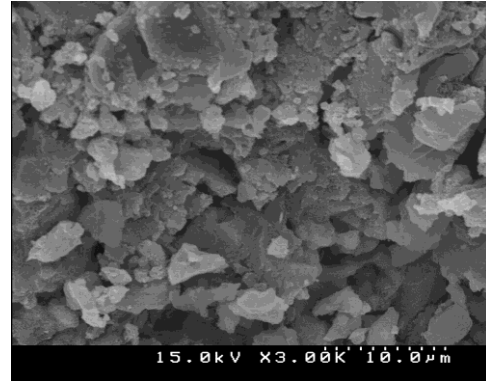
(c)

Fig. 2. Relation between the curing condition and compressive strength of AAS mortar with different Na<sub>2</sub>O concentrations: (a) 4 %, (b) 6 %, and (c) 8 %.

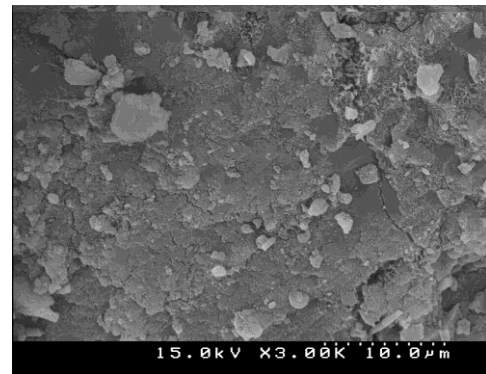
### C. Microstructure analysis

Scanning electron microscopy (SEM) was used in order to access the microstructure of AAS mortar under effect of different curing methods and Na<sub>2</sub>O concentrations. Figure 3 depicts the surface of mortar fragment A4 cured by different three regimes. As a consequence of the ambient curing

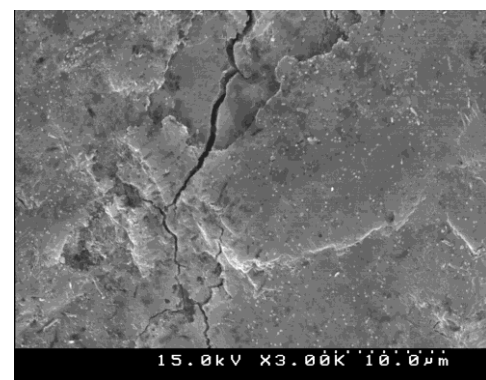
condition, AAS mortar had a porous structure at the age of 28 days (Figure 3a). It is noted that the mortar structure became denser with using the heat treatment methods (Figure 3b and 3c). This can be used to explain for the strength improvement in AAS mortar, stemming from the H and W method. As curing in the saturated lime water at 80 °C, the mortar structure was observed to be more well packed than that cured in a dry oven.



(a)



(b)

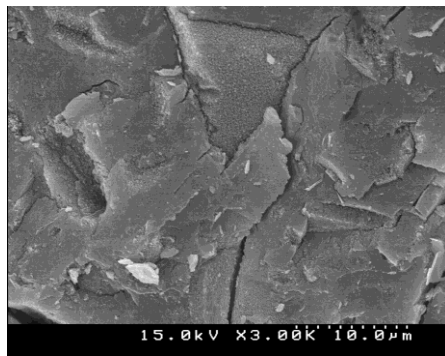


(c)

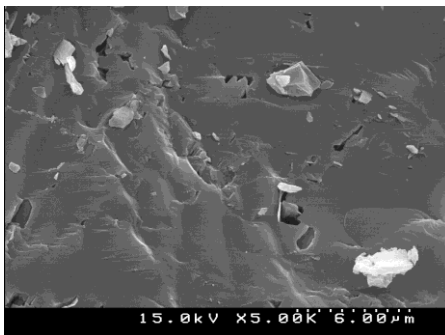
Fig. 3. SEM images of fracture surface of A4 at 28 days cured by different methods: (a) A, (b) H, and (c) W.

Figure 4 exposes the micrograph of the ambient temperature cured AAS mortar matrix corresponding to the activator with Na<sub>2</sub>O concentration of 4, 6, and 8 %, respectively. As shown in Figure 4, the mortar structure was

more compact owing to higher concentration of  $\text{Na}_2\text{O}$ . Increasing  $\text{Na}_2\text{O}$  dosage led to accelerate the alkali – activation kinetic of slag. As a result, more produced C-S-H can lead to the void filling in AAS mortar structure, resulting in the promotion in the compressive strength.



(a)



(b)

Fig. 4. SEM images of fracture surface of A4 at 28 days with the  $\text{Na}_2\text{O}$  concentration of 6 % (a), and 8 % (b).

The microstructure of A4 specimen at 120 days curing age is shown in Figure 5. As compared with Figure 3, the mortar structure was found to be more compact with longer curing age from 28 days to 120 days, which is responsible for the strength gain at 120 days. The result, as shown in Figure 3 and 5, indicate that there was a remarkable transformation from the porous structure at 28 days to relatively dense structure of AAS mortar at 120 days. This matrix alteration can prove for the significant rise in the mechanical strength of the ambient cured specimen with longer curing period. In contrast, the mortar matrix using the H curing method appeared with cracks and grain detachment, which resulted in the low strength improvement with curing age.

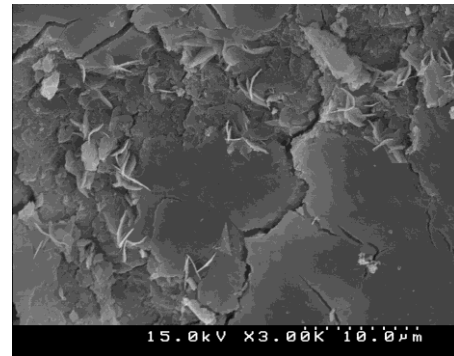
#### IV. CONCLUSION

Based on the experimental results and discussion above, the following conclusions can be drawn:

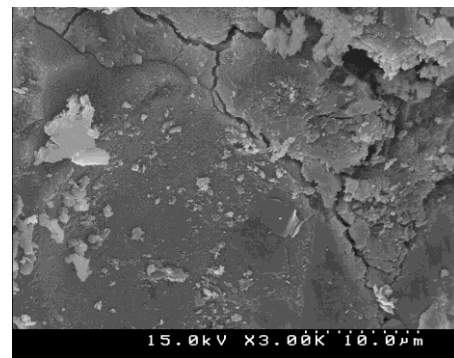
Increasing the  $\text{Na}_2\text{O}$  concentration of the alkaline solution can cause the compressive strength promotion of AAS mortar. However, this strength gain of mortar cured with heat treatment (H and W method) reduced when raising the  $\text{Na}_2\text{O}$  concentration from 6 to 8 %. This reduction in strength gain was negligible in the specimens using the ambient condition for curing (A method).

The curing condition in saturated limewater at 80 °C can effectively enhance the compressive strength of AAS mortar.

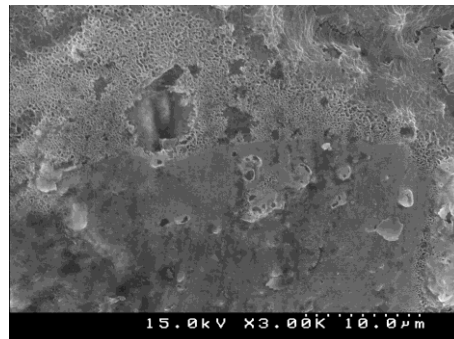
In addition, the specimen cured in the ambient regime can attain the high strength at long – term age despite its low strength at early age. The strength gain due to heat treatment was found to be higher with lower concentration of  $\text{Na}_2\text{O}$ .



(a)



(b)



(c)

Fig. 5. SEM micrographs of fracture surface of A4 at 120 days with different curing method: (a) A, (b) H, and (c) W.

#### REFERENCES

- [1] K. Kermeli et al., "The scope for better industry representation in long-term energy models: Modeling the cement industry," *Appl. Energy*, vol. 240, no. January, pp. 964–985, 2019, doi: 10.1016/j.apenergy.2019.01.252.
- [2] Roy D M, "Alkali activated cements, opportunities and challenges," *Cem. Concr. Res.*, vol. 29, pp. 249–254, 1999.
- [3] F. Puertas, T. Amat, A. Fernández-Jiménez, and T. Vázquez, "Mechanical and durable behaviour of alkaline cement mortars reinforced with polypropylene fibres," *Cem. Concr. Res.*, vol. 33, no. 12, pp. 2031–2036, 2003, doi: 10.1016/S0008-8846(03)00222-9.
- [4] J. E. Oh, P. J. M. Monteiro, S. S. Jun, S. Choi, and S. M. Clark, "The evolution of strength and crystalline phases for alkali-activated ground blast furnace slag and fly ash-based geopolymers," *Cem. Concr. Res.*, vol. 40, no. 2, pp. 189–196, 2010, doi: 10.1016/j.cemconres.2009.10.010.



- [5] P. K. Mehta and P. J. M. Monteiro, *Concrete: microstructure, properties, and materials*. 2006.
- [6] T. Bakharev, J. G. Sanjayan, and Y. B. Cheng, "Alkali activation of Australian slag cements," *Cem. Concr. Res.*, vol. 29, no. 1, pp. 113–120, 1999, doi: 10.1016/S0008-8846(98)00170-7.
- [7] A. Fernández-Jiménez, J. G. Palomo, and F. Puertas, "Alkali-Activated Slag Mortars Mechanical Strength Behavior," *Cem. Concr. Res.*, vol. 29, pp. 1313–1321, 1999.
- [8] T. Bakharev, J. G. Sanjayan, and Y. B. Cheng, "Sulfate attack on alkali-activated slag concrete," *Cem. Concr. Res.*, vol. 32, no. 2, pp. 211–216, 2002, doi: 10.1016/S0008-8846(01)00659-7.
- [9] T. Bakharev, J. G. Sanjayan, and Y. B. Cheng, "Resistance of alkali-activated slag concrete to acid attack," *Cem. Concr. Res.*, vol. 33, no. 10, pp. 1607–1611, 2003, doi: 10.1016/S0008-8846(03)00125-X.
- [10] H. T. Türker, M. Balçikanlı, I. H. Durmuş, E. Özbay, and M. Erdemir, "Microstructural alteration of alkali activated slag mortars depend on exposed high temperature level," *Constr. Build. Mater.*, vol. 104, pp. 169–180, 2016, doi: 10.1016/j.conbuildmat.2015.12.070.
- [11] M. Guerrieri, J. Sanjayan, and F. Collins, "Residual compressive behavior of alkali-activated concrete exposed to elevated temperatures," *Fire Mater.*, vol. 33, pp. 51–62, 2009.
- [12] F. Pacheco-Torgal, J. Castro-Gomes, and S. Jalali, "Alkali-activated binders: A review. Part 1. Historical background, terminology, reaction mechanisms and hydration products," *Constr. Build. Mater.*, vol. 22, no. 7, pp. 1305–1314, 2008, doi: 10.1016/j.conbuildmat.2007.10.015.
- [13] S. D. Wang and K. L. Scrivener, "Hydration products of alkali activated slag cement," *Cem. Concr. Res.*, vol. 25, no. 3, pp. 561–571, 1995, doi: 10.1016/0008-8846(95)00045-E.
- [14] B. S. Gebregziabihier, R. J. Thomas, and S. Peethamparan, "Temperature and activator effect on early-age reaction kinetics of alkali-activated slag binders," *Constr. Build. Mater.*, vol. 113, pp. 783–793, 2016, doi: 10.1016/j.conbuildmat.2016.03.098.
- [15] S. Aydin and B. Baradan, "Effect of activator type and content on properties of alkali-activated slag mortars," *Compos. Part B Eng.*, vol. 57, pp. 166–172, 2014, doi: 10.1016/j.compositesb.2013.10.001.
- [16] D. Krizan and B. Zivanovic, "Effects of dosage and modulus of water glass on early hydration of alkali-slag cements," *Cem. Concr. Res.*, vol. 32, no. 8, pp. 1181–1188, 2002, doi: 10.1016/S0008-8846(01)00717-7.
- [17] M. Chi, "Effects of dosage of alkali-activated solution and curing conditions on the properties and durability of alkali-activated slag concrete," *Constr. Build. Mater.*, vol. 35, pp. 240–245, 2012, doi: 10.1016/j.conbuildmat.2012.04.005.
- [18] S. Aydin and B. Baradan, "Mechanical and microstructural properties of heat cured alkali-activated slag mortars," *Mater. Des.*, vol. 35, pp. 374–383, 2012, doi: 10.1016/j.matdes.2011.10.005.
- [19] E. Altan and S. T. Erdoğan, "Alkali activation of a slag at ambient and elevated temperatures," *Cem. Concr. Compos.*, vol. 34, no. 2, pp. 131–139, 2012, doi: 10.1016/j.cemconcomp.2011.08.003.
- [20] E. Deir, B. S. Gebregziabihier, and S. Peethamparan, "Influence of starting material on the early age hydration kinetics, microstructure and composition of binding gel in alkali activated binder systems," *Cem. Concr. Compos.*, vol. 48, pp. 108–117, 2014, doi: 10.1016/j.cemconcomp.2013.11.010.
- [21] A. R. Brough and A. Atkinson, "Sodium silicate-based, alkali-activated slag mortars - Part I. Strength, hydration and microstructure," *Cem. Concr. Res.*, vol. 32, no. 6, pp. 865–879, 2002, doi: 10.1016/S0008-8846(02)00717-2.
- [22] B. S. Gebregziabihier, R. Thomas, and S. Peethamparan, "Very early-age reaction kinetics and microstructural development in alkali-activated slag," *Cem. Concr. Compos.*, vol. 55, pp. 91–102, 2015, doi: 10.1016/j.cemconcomp.2014.09.001.
- [23] T. Bakharev, J. G. Sanjayan, and Y. B. Cheng, "Effect of elevated temperature curing on properties of alkali-activated slag concrete," *Cem. Concr. Res.*, vol. 30, no. 9, pp. 1367–1374, 2000, doi: 10.1016/S0008-8846(00)00349-5.
- [24] ASTM C109-02, "Standard Test Method for Compressive Strength of Hydraulic Cement Mortars," in *Annual book of ASTM standards*, USA, 2002.
- [25] C. Bilim and C. D. Ati, "Alkali activation of mortars containing different replacement levels of ground granulated blast furnace slag," *Constr. Build. Mater.*, vol. 28, no. 1, pp. 708–712, 2012, doi: 10.1016/j.conbuildmat.2011.10.018.
- [26] C. Duran Atiş, C. Bilim, Ö. Çelik, and O. Karahan, "Influence of activator on the strength and drying shrinkage of alkali-activated slag mortar," *Constr. Build. Mater.*, vol. 23, no. 1, pp. 548–555, 2009, doi: 10.1016/j.conbuildmat.2007.10.011.

# Design and Control of a 4-bar-transmission 2-DOF Robot

Pham Tan Phat

*Ho Chi Minh City University of  
Technology and Education (HCMUTE)*

Ho Chi Minh City, Vietnam  
halophat290799@gmail.com

Bui Manh Huy

*Ho Chi Minh City University of  
Technology and Education (HCMUTE)*

Ho Chi Minh City, Vietnam  
buimanhhuy13@gmail.com

Tran Hoai Nam

*Ho Chi Minh City University of  
Technology and Education (HCMUTE)*

Ho Chi Minh City, Vietnam  
namtran.spk@gmail.com

Huynh Vinh Nghi

*Ho Chi Minh City University of Technology and Education  
(HCMUTE)*

Ho Chi Minh City, Vietnam  
vinh.nghi.810@gmail.com

Dang Xuan Ba\*

*Ho Chi Minh City University of Technology and Education  
(HCMUTE)*

Ho Chi Minh City, Vietnam  
badx@hcmute.edu.vn

**Abstract**—The paper presents design and position control of 2-degree-of-freedom robot. To satisfy special requirements of massless distribution and fast response, transmission mechanism of the robot is designed using a 4-bar structure. To effectively support the high-precision control process, an intensive inverse kinematic algorithm is studied. A neural-network-based proportional-integral-derivative (PID) controller is then developed using a nonlinear learning law. Effectiveness of the robot is confirmed by simulation and real-time experiments.

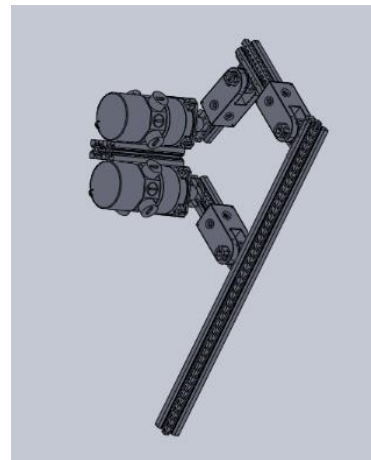
**Keywords**—Leg robot, 4-bar robot, Robot kinematics, PID controller, Intelligent Controller

## I. INTRODUCTION

Nowadays, with the rapid development of science and technology, the robots have significantly supported people in many fields such as industry, healthcare, or rescue missions. Among the robotics sciences, leg robots have increasingly attracted the researchers' attention both in design and control aspects. Not similar to industrial robots, the leg robots need distinct design and robust controller [1]. Using legs will indeed make robots more flexible than wheels in the complex terrain. The fastest quadruped robot so far is the Cheetah-Cub robot [4]. The robot uses rods, instead of using the belt, and shock absorbers in transmission design which help robot easier to reduce impact force from the ground [5]. However, using of multiple rods could increase the size of the robot legs [2].

To overcome this drawback, a research team has come up with a new two-segment robot [7]. The structure of this robot is using less rods, but it is heavier than the Cheetah-Cub [8]. One of big advantages is the ability of recovering itself from external effects thanks to a special shock absorber set up at the knee joint. Although this robot could result in a strong balance, it might yield the poor performance in flat terrain. A new robot has been designed with the combination of wheel and leg structures. This robot is capable of moving by wheels and feet [9]. By employing wheels as feet, it provided a faster speed in perfect terrains and the shock absorbers could well suppress unpredictable impacts in complex terrains [10]. Nevertheless, this robot also has some disadvantages such as: it is difficult to change the moving direction at sharp corners.

To accomplish control objectives of the robots in real-life missions, normally simple proportional-integral-derivative (PID) controllers are favored [11],[12]. If the proper control could be found, the high control outcome was obtained [13], [15]. A number of researches have been studied to improve the performance of the PID controllers using intelligent approaches such as evolutionary optimization and fuzzy logic [14]. The methods exhibited promising control results thanks to using both online and offline learning features [16]. The off-line control one could flexibly select the proper PID parameters based on the system overshoot,



a) The Solidworks model



b) The realistic model

Fig. 1. The studied leg robot

settling time and steady-state error, while the on-line one would adopt the operating control errors to adjust fuzzy logic parameters to re-optimize the system, improving the system quality significantly. However, the tuning methodology of fuzzy logic controllers are mostly based on the operator's experience [17]. Another series of the intelligent control category was based on the biological properties of animals in which a genetic algorithm was combined with a bacterial foraging method to simulate natural optimization processes such as hybridization, reproduction, mutation, natural selection, etc. [18]. This evolutionary could deliver the most optimal solution. That the solving process requires a large number of samples and takes a long running time limits its application. Recently, tuning PID control parameters using neural networks has become effective approaches attracting many contributions [19], [20]. The conventional PID one itself is a robust controller [21]. Integrating the learning ability to the controllers makes it flexible to the working environment. Lack of intensive consideration of learning rules in steady-state time could make the system instable in a long time used [3], [22], [23].

In this paper, research results of a leg robot are reported including mechanical design, kinematics computation and an intelligent control algorithm. A direct transmission mechanism is employed based on a four-link structure that results in fast responses and a massless-like design. An associated kinematic computation is presented to provide necessary materials for the control phase. A robust adaptive controller is then designed to maintain the good performance in variation working conditions. The learning law of the controller is activated by nonlinear-segment excitation signals that could ensure stability of the closed-loop system. The entire robot was fabricated and carefully tested achieving promising results.

The rest of this paper is organized as follows. The robot configuration and its kinematics are presented in Section II. The neural network controller proposed is derived in Section III. Validation results on simulation environment and real-time experiments are discussed in Section IV, and the paper

is then concluded in Section V.

## II. ROBOT KINEMATICS

The studied robot was designed with two degree of freedoms in which its transmission mechanism is a four-link structure. The robot model designed in Solidworks software and a fabricated apparatus are shown in **Fig. 1**. The key feature of this design is to make sure the center of mass distributed in the hip joint and is also able to yield a large range of motion. Then, the system needs solution of the inverse kinematics to perform high accuracy control tasks.

Based on the robot configuration, which is simply drawn in **Fig. 2**, the desired joint angles  $(\theta_{1d}, \theta_{2d})$  are computed from a given end-effector position  $(x_{EE}, y_{EE})$ , as follows:

$$\begin{cases} d = (x^2 + y^2 + L_1^2 - L_2^2) / (2L_1) \\ \theta_1 = \text{atan2}(y, x) + \text{atan2}(-\sqrt{x^2 + y^2 - d^2}, d) \\ \theta_2 = \text{atan2}(x - L_1 \sin(\theta_1), x - L_1 \cos(\theta_1)) - \theta_1 \end{cases} \quad (1)$$

In this design, the first joint  $(\theta_1)$  is actuated by a DC motor placed at A position in **Fig. 2**, while motion of the second joint  $(\theta_2)$  is transferred from another DC motor placed in C position. To realize the aforecomputed second joint angle  $(\theta_{2d})$ , a second-level inverse-kinematics computation is applied as follows:

$$\begin{cases} b_1 = \sqrt{a_3^2 + a_4^2 - 2a_3a_4 \cos \theta_2} \\ \beta_1 = \text{acos}\left(\frac{a_4^2 + b_1^2 - a_3^2}{2a_4b_1}\right) \\ \begin{cases} \beta_2 = \frac{\pi}{2} + \theta_1 - \beta_1 \\ b_2 = \sqrt{a_5^2 + b_1^2 - 2a_5b_1 \cos \beta_2} \\ \gamma_1 = \text{acos}\left(\frac{a_5^2 + b_2^2 - b_1^2}{2a_5b_2}\right) \end{cases} & (0 < \beta_1 < \pi; 0 < \beta_2 < \pi) \\ \begin{cases} \gamma_2 = \text{acos}\left(\frac{a_1^2 + b_2^2 - a_2^2}{2a_1b_2}\right) \\ \theta_3 = \gamma_1 + \gamma_2 - \frac{\pi}{2} \end{cases} & (0 < \gamma_1 < \pi; 0 < \gamma_2 < \pi) \end{cases} \quad (2)$$

where  $\theta_3, a_{i|i=1..5}, b_1, b_2, \gamma_1, \gamma_2$  are marked in detail in **Fig. 2**.

For controlling this robot, a simple PID controller could be adopted for an acceptable control result. Note that the robot leg would be working in a complex terrain that encompasses a lot of unpredictable factors waiting for destroying the system. As a sequence, designing an intelligent controller that could be adapted well with complicated environments becomes a very important requirement.

## III. INTELLIGENT CONTROLLER

Framework of the intelligent controller is based on a convention PID structure as expressed in Eq. (3).

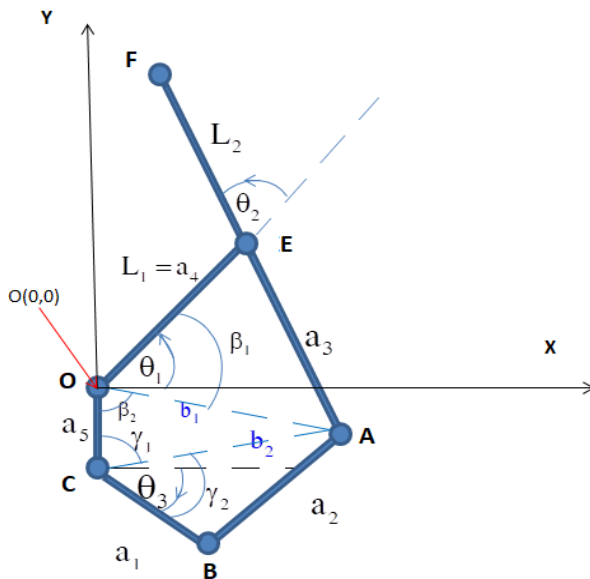


Fig. 2. Configuration of the studied leg robot

$$\boldsymbol{\tau} = \mathbf{K}_p \mathbf{e} + \mathbf{K}_D \dot{\mathbf{e}} + \mathbf{K}_I \int_0^T \mathbf{e} d\tau \quad (3)$$

where  $\mathbf{K}_p, \mathbf{K}_D, \mathbf{K}_I$  are PID control gains, and  $\mathbf{e} = \boldsymbol{\theta}_d - \boldsymbol{\theta}$  is the joint control error,  $\boldsymbol{\tau}$  are the motor torque or the control input.

In real-time control, one needs to tune the control gains for different working conditions. To support this feature, an automatic tuning mechanism is designed as follows:

$$\begin{cases} \dot{\mathbf{K}}_p = \beta_p \mathbf{e}^2 (\text{sgn}(\mathbf{e}) - \mathbf{e}_0) \\ \dot{\mathbf{K}}_I = \beta_I \sqrt{|\mathbf{e}|} (\text{sgn}(\mathbf{e}) - \mathbf{e}_0) \int_0^T \mathbf{e} d\tau \\ \dot{\mathbf{K}}_D = \beta_D \mathbf{e}^{2/3} (\text{sgn}(\mathbf{e}) - \mathbf{e}_0) \end{cases} \quad (4)$$

where  $\beta_{p,D,I}$  are learning rates and  $\mathbf{e}_0$  is a predefined steady-state error.

As noted in Eq. (4), the control gains are varied in nonlinear manners to force the control error go into a desired region regardless of unknown environments. The learning of The  $\mathbf{K}_p$  and  $\mathbf{K}_I$  gains is respectively designed for fast transient time and the best steady-state phase, while that of the  $\mathbf{K}_D$  acts as a trade-off between the two phases. The scheme of the proposed controller is shown in Fig. 3.

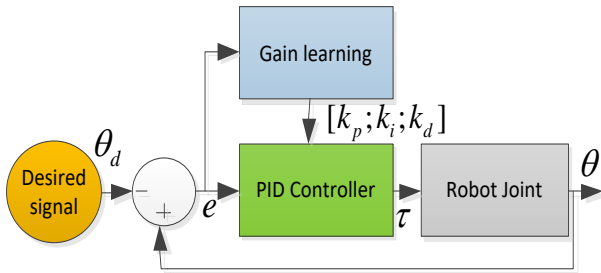


Fig. 3. Control scheme of the robot.

#### IV. VALIDATION

This section presents the testing results of the designed robot and the associated controller both in simulations and real-time experiments. To carefully express the performance of the proposed controller, a conventional PID controller was also implemented on the same system.

##### A. Simulation Results

Dynamics of the robot was first derived based on the real parameters of the fabricated robot [24]. Simulation results of the conventional and intelligent PID controllers are shown in Fig. 4 and Table 1. As seen in Fig. 4 a and b, two controllers could provide the excellent steady-state error, but the intelligent control method improved well the transient response without overshoot even in the same settling time. This superior property was the achievement of the learning law (4) that is demonstrated by the gain variation as depicted in Fig. 4 c. Here, the advantages of the proposed controller have been confirmed.

##### B. Real-time Experiments

Before conducting the real-time test, workspace of the robot was computed using the two-level forward kinematics

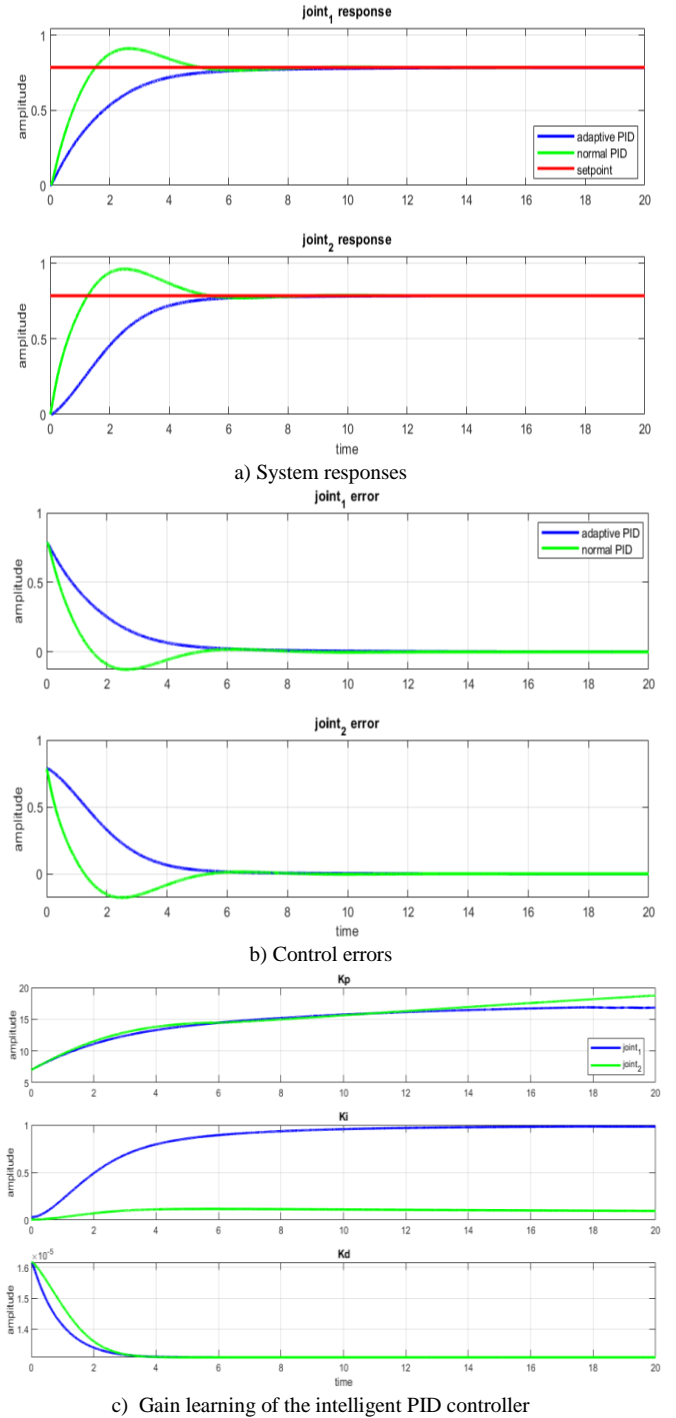


Fig. 4. Comparative control performance of the PID controllers in simulation.

TABLE I. COMPARISON OF STATISTIC PERFORMANCES OF THE TWO CONTROLLERS IN SIMULATION TESTS

Criterion	Joint1		Joint2	
	PID	A-PID	PID	A-PID
Overshoot (%)	15.92	0	22.68	0
Steady-state error (rad)	0.002	0.0015	0.0023	0.0018
Settling time (s)	4.29	4.84	4.58	4.58

derived in *Appendix A*. The calculation of the workspace could help the control process of the robot more secure. If we know in advance the workspace, we can easily control the robot tracking the desired trajectories, such as lines, trails, arc, or ellipses etc., without having any worry about mechanics. **Figure 5** shows the workspace of the real robot and the simulation result in drawing an arc.

The two controllers were deployed to perform a tracking control test that drawn the arc as illustrated in **Fig. 5**. The experimental results obtained are shown in **Fig. 6**. We can visually see the difference results of two methods. The conventional one gave the arc that was not smooth and had some vibration in a brief moment, as displayed in zoomed-in sub-figures, but after that it was quite stable for the rest. The main problem is that the control coefficients  $K_P$ ,  $K_I$  and  $K_D$  were fixed values during the whole working process. Such the constant gains could not maintain the control performance in different loading regions. Meanwhile, the intelligent one was almost stable in the whole process. Thanks to the gain learning mechanism, the control gains were self-adjusted and generated proper control signals which made the system working smoothly, even in the heavy-load region of the system. The effectiveness of the proposed system has been re-confirmed throughout this experiment. Further videos of the test could be found in Youtube: Video of the conventional PID: <https://youtu.be/h2RoPa6yYg8>, video of the intelligent PID controller: <https://youtu.be/nhyUEGLjNOo>.

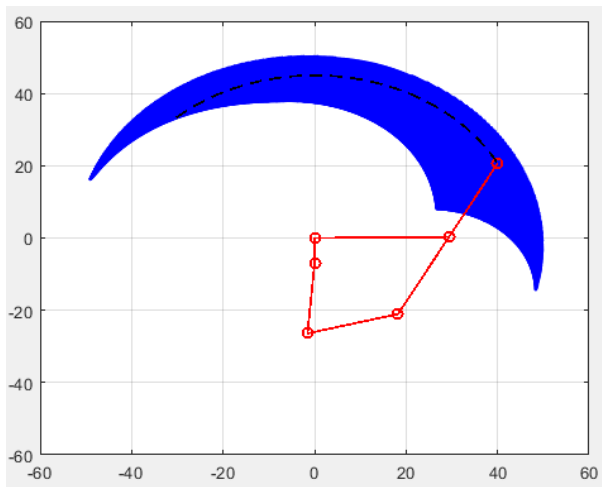


Fig. 5. Workspace and behavior of the real robot drawing an arc

## V. CONCLUSION

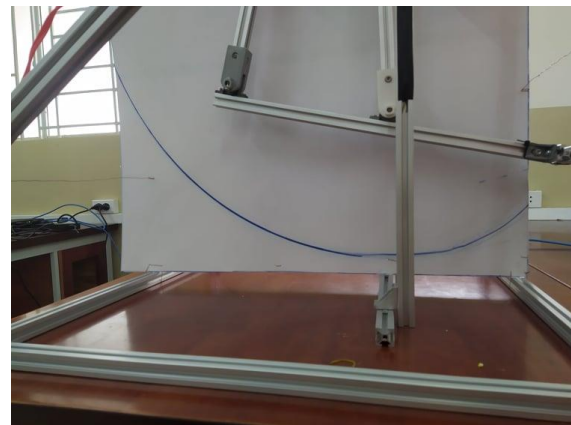
In this paper presents, a leg robot and an associated intelligent controller are designed for tracking control tasks. To satisfy a massless distribution and to provide fast responses, transmission mechanism of the robot is actuated using a 4-bar structure. To effectively support the high-precision control process, a two-level inverse kinematic algorithm is studied. A neural-network-based proportional-integral-derivative (PID) controller is then developed using a nonlinear-segment learning law. Effectiveness of the entire system is strongly confirmed by simulation and real-time experiments.



a) Performance of the conventional PID controller



b) Zoomed-in control result of the conventional PID controller



b) Performance of the intelligent PID controller



d) Zoomed-in control result of the intelligent PID controller

Fig. 6. Tracking real-time control results of the two controllers



#### APPENDIX A: TWO LEVELS OF FORWARD KINEMATICS OF THE 4-BAR ROBOT

Scenario of this computation is to find out the end-effector position from a given set of joint angles  $\theta_1$  and  $\theta_3$ . To this end, the kinematics could be obtained by computing the relatedly red edges drawn in **Fig. A. 1**. Note that the lengths of links ( $L_1, L_2, a_1, a_2, a_3, a_4, a_5$ ) are known. By considering the triangle **OCB**, it can result in as:

$$\begin{cases} OB = \sqrt{a_1^2 + a_5^2 - 2a_1a_5 \cos\left(\frac{\pi}{2} + \theta_3\right)} \\ COB = \arccos\left(\frac{a_5^2 + OB^2 - a_1^2}{2a_5 \cdot OB}\right) \end{cases} \quad (A1)$$

From the given  $\theta_3$  and **C** point, then we have:

$$\begin{cases} BOx = -a \tan 2(yB, xB) \\ BOE = BOx + \theta_1 \end{cases} \quad (A2)$$

The computation is continued by noting the Triangle **OEB** that provides:

$$\begin{cases} EB = \sqrt{a_4^2 + OB^2 - 2a_4 \cdot OB \cdot \cos(BOE)} \\ OEB = \arccos\left(\frac{a_4^2 + EB^2 - OB^2}{2a_4 \cdot EB}\right) \end{cases} \quad (A3)$$

Finally, the end-effector position is determined using the following relationship:

$$\begin{cases} AEB = \arccos\left(\frac{a_3^2 + EB^2 - a_2^2}{2a_3 \cdot EB}\right) \\ \theta_2 = OEB + AEB \\ x = L_1 \cos(\theta_1) + L_2 \cos(\theta_1 + \theta_2) \\ y = L_1 \sin(\theta_1) + L_2 \sin(\theta_1 + \theta_2) \end{cases} \quad (A4)$$

#### ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.01-2020.10.

#### REFERENCES

- [1] P. E. Hudson, S. A. Corr and A. M. Wilson, "High speed galloping in the cheetah (*acinonyx jubatus*) and the racing greyhound (*canis familiaris*): spatio-temporal and kinetic characteristics," *Journal of Experimental Biology*, vol. 215, no. 14, pp. 2425–2434, 2012.
- [2] H. Yeom, D. X. Ba, and J. B. Bae, "Design Principles and Validation of a Human-sized Quadruped Robot Leg for High Energy Efficiency," *J. K. Robotics Society*, vol. 13, no. 2, pp. 86–91, 2018.
- [3] D. X. Ba, H. Yeom, and J. B. Bae, "A Direct Robust Nonsingular Terminal Sliding Mode Controller based on an Adaptive Time-delay Estimator for Servomotor Rigid Robots," *Mechatronics*, May 2019.
- [4] K. Sreenath, H.-W. Park, I. Poulakakis, and J. W. Grizzle, "A compliant hybrid zero dynamics controller for stable, efficient and fast bipedal walking on MABEL," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1170–1193, August 2011.

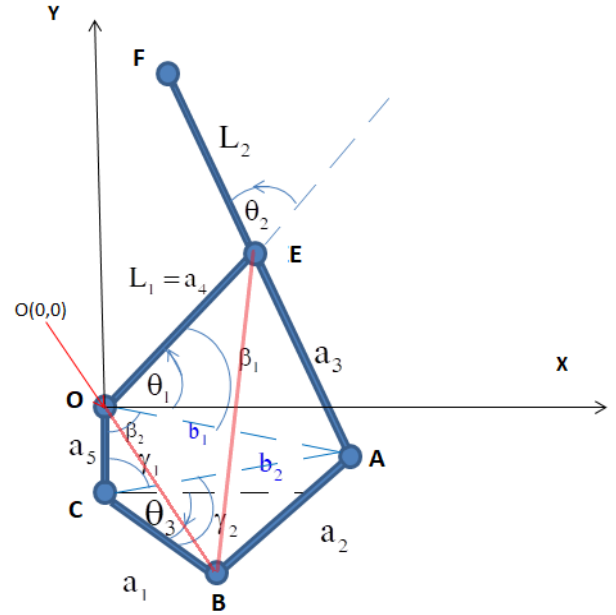


Fig. A. 1: The robot configuration associated with the important edges used for the forward kinematics.

- [5] K. Sreenath, H.-W. Park, I. Poulakakis, and J. W. Grizzle, "Embedding active force control within the compliant hybrid zero dynamics to achieve stable, fast running on MABEL," *International Journal of Robotics Research*, vol. 32, no. 3, pp. 324–345, March 2013.
- [6] S. Ruthishauser, *Cheetah – compliant quadruped robot*, Biologically Inspired Robotics Group, EPFL, 2008.
- [7] M. Hutter, C. Gehring, M. Bloesch, and et. al., "StarLETH: A compliant quadrupedal robot for fast, efficient, and versatile locomotion," *the 15th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines*, 2012.
- [8] R. Siegfried, *Effect of Leg Design on Locomotion Stability for Quadruped Robot*, WS 2014-2015, STI-SMT, Semester Project, 05.06.2015.
- [9] *Static Balancing of Wheeled-legged Hexapod Robots*, CICATA Instituto Politecnico Nacional – Unidad Queretaro 76090, Mexico, IRCCS Neuromed, 86077 Pozzilli, Italy DIMEG, University of Calabria, 87036 Cosenza, Italy 7, April, 2020.
- [10] K. Hashimoto, et al., "Realization by biped leg-wheeled robot of biped walking and wheel-driven locomotion," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Barcelona, Spain, 2005.
- [11] G. Bledt, M. J. Powell, B. Katz, F. D. Carlo, P. W. Wensing, and S. Kim, "MIT Cheetah 3: Design and Control of a Robust, Dynamic Quadruped Robot," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018.
- [12] P. M. Wensing, A. Wang, S. Seok, A. Otten, J. Lang, and S. Kim, "Proprioceptive Actuator Design in the MIT Cheetah: Impact Mitigation and High-Bandwidth Physical Interaction for Dynamic Legged Robots," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 509–522, 2017.
- [13] H. W. Park, S. Park, and S. Kim, "Variable-speed quadrupedal bounding using impulse planning: Untethered high-speed 3D Running of MIT Cheetah 2," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, USA, 2015.
- [14] K. Astrom and K. Hagglund, *PID Controllers: theory, design and tuning*. USA: ISA Press, 1995.
- [15] G. Z. Tan, Q. D. Zeng, and W. B. Li, "Intelligent PID controller based on ant system algorithm and fuzzy inference and its application to bionic artificial leg," *Journal of Central South University of Technology*, vol. 11, pp. 316–322, 2004.
- [16] C. F. Juang and Y. C. Chang, "Evolutionary-Group-Based Particle-Swarm-Optimized Fuzzy Controller With Application to Mobile-

- Robot Navigation in Unknown Environments,” *IEEE Trans. Fuzzy Systems*, vol. 19, no. 02, pp. 379-392, 2011.
- [17] M. Cuciates, D. L. Moreno, A. Bugarin, and S. Barro, “Design of a fuzzy controller in mobile robotics using genetic algorithm,” *Applied Soft Computing*, vol. 7, no. 2, pp. 540-546, 2007.
- [18] D. H. Kim and J. H. Cho, “A Biological Inspired Intelligent PID Controller Tuning for AVR Systems,” *International Journal of Control, Automation, and Systems*, vol. 4, no. 5, pp. 624 – 636, 2006.
- [19] M. J. Neath, A. K. Swain, U. K. Madawala, and D. J. Thrimawithana, “An Optimal PID Controller for a Bidirectional Inductive Power Transfer System Using Multiobjective Genetic Algorithm,” *IEEE Trans. Power Electronics*, vol. 19, no. 3, pp. 1523-1531, 2014.
- [20] T. D. C. Thanh and K. K. Ahn, “Nonlinear PID control to improve the control performance of 2 axes pneumatic artificial muscle manipulator using neural network,” *Mechatronics*, 2006.
- [21] J. Ye, “Adaptive control of nonlinear PID-based analog neural networks for a nonholonomic mobile robot,” *Neurocomputing*, 2008.
- [22] P. Roco, “Stability of PID control for industrial robot arms,” *IEEE Trans. Robot. Automation*, vol. 12, no. 4, pp. 606-614, 1996.
- [23] H. V. A. Truong, D. T. Tran, and K. K. Ahn, “A neural network based sliding mode control for tracking performance with parameters variation of a 3-dof manipulator,” *Applied Sciences*, 2019.
- [24] J. J. Craig, “Manipulator dynamics” in *Introduction to Robotics: Mechanics and Control*, 3rd ed, Pearson Prentice Hall, USA, 2005, pp. 165-200

# The Morphological Characteristics and Physical Properties of Porous Corn Starch Hydrolyzed by Mixture of $\alpha$ -Amylase and Glucoamylase

Dang My Duyen Nguyen

Department of Food Technology  
Ho Chi Minh City University of Education and Technology  
Ho Chi Minh City, Vietnam  
myduyen@hcmute.edu.vn

Thanh Tung Pham

Department of Food Technology  
Ho Chi Minh City University of Education and Technology  
Ho Chi Minh City, Vietnam  
tungpt@hcmute.edu.vn

**Abstract**—This research was conducted to evaluate the effect of  $\alpha$ -amylase and glucoamylase mixture at various hydrolysis time (0, 4, 6, 8 and 12 hours) to morphological characteristics and physical properties of porous corn starch granules. Scanning Electron Microscopy (SEM) showed the number and size of holes on the surface of starch granules increased with hydrolysis time; moreover, starch granule structure was broken at 8 and 12 hours. The hydrolyzed level of starch granules was increased gradually with reaction time; the highest hydrolysis value was 63.534% in 12 hours of reaction. The results indicated that porous starch had higher adsorption capacity than native starch which was showed in higher oil and water holding ability. Additionally, swelling and solubility of porous corn starch were increased with the prolonging of hydrolysis time and higher than that of native starch samples. Color value tended to increase with increasing hydrolysis time; however, the difference was very small. The results also indicated that viscosity and transmittance of the hydrolyzed starch sample decreased as the hydrolysis time increased as well as lower than native starch. Thus, porous corn starch could be a potential ingredient in the food field.

**Keywords**—porous corn starch; native starch; hydrolysis;  $\alpha$ -amylase; glucoamylase

## I. INTRODUCTION

Modified starch by the enzyme to create porous starch is receiving a lot of attention due to a number of outstanding advantages in terms of safety, low cost, and has a structure with function as expected [1]. Due to its large specific surface area of structure, semi-hydrolyzed corn starch is being studied and applied in food, medicine, cosmetics, and some other fields. In recent times, structural changes of hydrolyzed corn starch by enzymes  $\alpha$ -amylase and amyloglucosidase have also been investigated by [2, 3]. Moreover, Dura [4] also compares the changes of properties between hydrolyzed corn starch with enzyme  $\alpha$ -amylase and amyloglucosidase. However, recent researches have not evaluated physicochemical properties as well as technological features of corn starch when hydrolyzing with a mixture of enzyme  $\alpha$ -amylase and amyloglucosidase. Therefore, the purpose of this study is to investigate changes in starch morphology and some properties of corn starch when hydrating with a mixture of enzyme  $\alpha$ -amylase and amyloglucosidase.

## II. MATERIALS AND METHODS

### A. Materials

Maize starch powder of ROQUETTE Riddhi Siddhi (India) was used in this study. Enzyme  $\alpha$ -amylase (Spezyme® Alpha-amylase) of Dupont and enzyme amyloglucosidase (Dextrozyme® E) of Novozymes were used for further experiments.

### B. Methods

#### Determining enzyme activity of $\alpha$ -amylase and amyloglucosidase

Enzyme activity unit (U) was defined as the amount of enzyme required to release 1  $\mu$ mol product from substrate in 1 minute under standard test conditions. In this study, the enzyme activity of  $\alpha$ -amylase and amyloglucosidase was measured by following Okwuenu [5].

#### Semi-hydrolyzed corn starch preparation

Method of preparing semi-hydrolyzed corn starch samples was performed according to Zhang [2] and modified to suit with experimental conditions. Citric acid-Disodium hydrogen phosphate buffer solution (pH 5.5) and corn starch with a ratio of  $V_{\text{Buffer}} / M_{\text{Starch}} = 8/1$  (mL/g) were prepared in the flask. Mixture of enzyme  $\alpha$ -amylase and amyloglucosidase with ratio of  $V_{\alpha\text{-amylase}} / V_{\text{Amyloglucosidase}} = 16/1$  (mL/mL) was added to starch suspension according to ratio of  $V_{\text{Enzyme}} / M_{\text{Starch}} = 0.03/1$  (mL/g). Starch hydrolysis was performed in thermostatic shaker (Edmund Buhler, Germany) at 50°C with shaking speed was 120 rpm. Symbols T0, T4, T6, T8, T12 was used for non-hydrolyzed corn starch sample, corn starch was hydrolyzed in 4, 6, 8 and 12 hours. After hydrolysis, sodium hydroxide 0.1M was added until pH increased to 10 to inactivate enzymes and stop hydrolysis reaction. Corn starch after hydrolysis was cleaned by washing with distilled water and centrifuged at  $2400 \times g$  for 10 minutes. This process was repeated 3 times. After that, corn starch was dried by vacuum drying device VO400 (Mettler, Germany) at 50°C, pressure of 50 bar in 8 hours. Starch after drying was finely ground, sifted (0.1 mm sieve), and stored in PE vacuum bag.

#### Starch granule structure evaluation

Morphological characteristics of native and semi-hydrolyzed starch were observed by scanning electron microscopy JSM 7401F (JEOL, Tokyo, Japan).

### Determining degree of hydrolysis

Corn starch hydrolysis level was determined by the following method of Zhang [2] with some adjustments. Tubes containing 10 mL of sample after hydrolysis were centrifuged at  $7000 \times g$  for 15 minutes. The supernatant fluid was carried away to determine total sugar content after hydrolysis by Phenol- sulfuric acid method (Dubois et al. 1956); and determined at a wavelength of 490 nm by using Halo VIS 20 Spectrophotometer (Dynamica, Switzerland). Degree of hydrolysis was determined by formula:

$$DH(\%) = \frac{M}{M_T} \times 100$$

where:  $M_T$  is the amount of starch before hydrolysis (g)

$M_S$  is the amount of hydrolyzed starch (g), which was determined by total sugar content produced and converted to the amount of hydrolyzed starch by multiplying with a coefficient of 0.9 [6].

### Determining Water holding capacity and Oil holding capacity

Oil holding capacity (OHC) was measured by the method of Sarangapani [7]. Tubes containing 0.1 g of starch and 1.0 mL of walnut oil then mixture was shaken well for 30 minutes by the shaker. After that, the centrifuge tube was centrifuged at  $3000 \times g$  for 10 minutes.

Water holding capacity (WHC) was determined by Kim [8]. Tubes containing 5% w/v suspension of starch powder were shaken for 10 minutes by the shaker. Then tubes were centrifuged at  $1000 \times g$  for 15 minutes. The supernatant fluid was removed and the weight of remaining substance was determined. Oil holding capacity and Water holding capacity were determined by following formula:

$$OHC(g/g) = \frac{W}{W_i}$$

$$WHC(g/g) = \frac{W}{W_i}$$

where:  $W_r$  is the weight of remaining substance after centrifuging (g)

$W_i$  is the weight of sample before centrifuging (g)

### Determining solubility and swelling power

The solubility, swelling power of native and semi-hydrolyzed starches were determined by Amini [9]. Tubes containing 1% w/v starch suspension were heated in water bath at  $60^\circ\text{C}$  for 30 minutes. Shaker was used continuously every 1 minute for 5 times. The tube was then cooled to ambient temperature and centrifuged at  $2400 \times g$  for 20 minutes. After centrifuging, the residue was weighed to determine mass while the supernatant was transferred to petri dish and dried at  $105^\circ\text{C}$  to constant weight to determine amount of solid present in solution. Solubility

(SB) and swelling power (SP) were calculated by following formulas:

$$SB(\%) = \frac{W}{W_i} \times 100$$

$$SP(g/g) = \frac{W_r}{W_i(1-SB)}$$

where:  $W_i$  is the weight of initial starch sample (g)

$W_s$  is the weight of solid remaining in supernatant after drying (g).

$W_r$  is the weight of residue in starch suspension after centrifuging (g)

### Determining viscosity

Viscosity of native starch and semi-hydrolyzed starches were measured by the method of Bello-Pérez [10] with some modification. Starch suspension of 6% w/v was boiled in water for 30 minutes. After that, starch sample was cooled down to the temperature of  $50^\circ\text{C}$  and kept thermally stable at  $50^\circ\text{C}$  in water bath (Mettler, Germany). The viscosity was determined by the Brookfield viscometer DV-II + Pro (USA). Measurement was repeated 3 times.

### Determining transmittance

Transmittance of native and semi-hydrolyzed starches was conducted by the method of Wani [11]. The starch suspension (1% w/v) after boiling for 30 minutes is cooled to room temperature. Starch samples were stored for 5 days at  $4^\circ\text{C}$  and measured total transmittance at 640 nm every 24 hours by using Halo VIS 20 Spectrophotometer (Dynamica, Switzerland).

### Determining colorimetry

Colorimetry was evaluated by ColorFlex EZ colorimeter (HunterLab, USA). The  $L^*$  value represents brightness that ranges from 0 (black) to 100 (white). Values  $a^*$  and  $b^*$  represent ranges from -a (green) to +a (red) and from -b (blue) to +b (yellow), respectively [12]. The whiteness of starch in CIE  $L^* a^* b^*$  color space was calculated using formula [13]:

$$WI = 100 - \sqrt{(100 - L^*)^2 + a^{*2} + b^{*2}}$$

The difference in color  $\Delta E$  was calculated by the formula:

$$\Delta E = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2}$$

where:  $\Delta L = L^*_{\text{sample}} - L^*_{\text{Control}}$   
 $\Delta a = a^*_{\text{sample}} - a^*_{\text{Control}}$   
 $\Delta b = b^*_{\text{sample}} - b^*_{\text{Control}}$

### III. RESULTS AND DISCUSSIONS

#### Morphological characteristics of starch granule

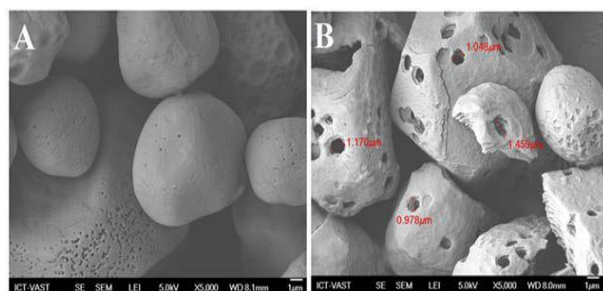


Figure 1. Morphology of native corn starch (A) and porous corn starch (B)

SEM results showed that native corn starch had spherical and polygon shape (Fig. 1-A). The surface of native corn starch was relatively smooth and had holes randomly distributed in clusters with different quantities in each granule. Surface structure of hydrolyzed corn starch was different while morphology was almost unchanged (Fig. 1-B). Hydrolyzed starch granules became rough and granule holes had wider diameter (about 0.2 to 1.7  $\mu\text{m}$ ) than native starch granules. This result was similar to researches of Dura and Diop [4, 14]. The reason for surface structure changing may be due to holes of native corn starch granules. These holes were the starting point for small grooves connecting surface to starch granule center [15]. And these grooves were mostly amorphous and easily hydrolyzed by enzymes [16]. Hydrolysis process expanded diameter of grooves and holes in the surface of starch granules [17]. The model of corn starch hydrolyzed using a mixture of enzyme  $\alpha$ -amylase and amyloglucosidase as above had been studied by Zhang, Helbert, Dhital [2, 18, 19] and had similar result.

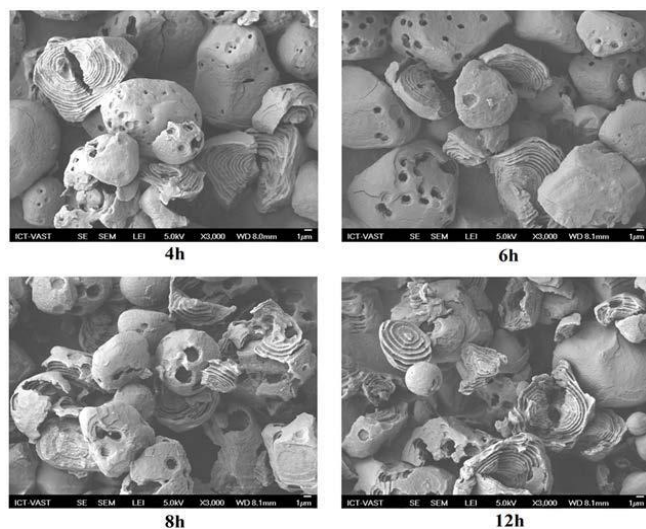


Figure 2. Morphology of corn starch granules in different hydrolysis times

SEM results at different times showed that hydrolysis time significantly affected to starch granule surface (Fig. 2). Some holes on the surface of starch granules when hydrolyzing would be enlarged and formed small holes. The number of formed small holes depends on the hydrolysis time. In T4 sample (4 hours of hydrolysis), holes on the surface appeared in small quantities. After 4 hours of hydrolysis, holes appeared

more and diameter also increased. When hydrolysis time increased to 8 hours and 12 hours, hole diameter was greatly expanded and some of the starch granules were broken down due to excessive hydrolysis. The above results were similar to those previously study by [20]

#### Effect of time on hydrolysis degree of starch granules

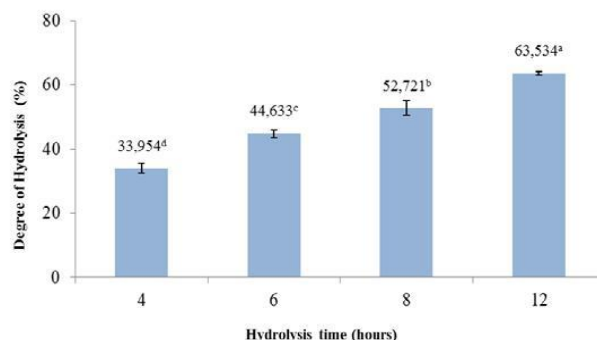


Figure 3. Corn starch hydrolysis degree at different time

Results indicated the degree of hydrolysis increased with time of hydrolysis (Fig. 3). This may be due to the surface and inner structure of starch granules becoming less tight during hydrolysis. Therefore, it was easier for the enzyme to penetrate deep into starch center and cut hydrogen bonds in the amorphous region [21]. These results were similar to research of Helbert [18].

#### The effect of hydrolysis time on oil holding capacity and water holding capacity

The result of Table 1 showed that OHC and WHC of hydrolyzed starch were higher than native starch. This may be due to the hydrolysis process by enzyme amylase that increased quantity, volume, pore size and grooves inside the starch granules [22]. This would increase the specific surface area of starch granule after hydrolysis thus helping starch granules to hold oil and water better [23]. Zhang [2] also reported that a change in structure during hydrolysis would significantly affect the ability of starch to retain oil and water. It also showed that the ability of starch to hold oil and water increased with hydrolysis time, but to a certain level it would be reduced. The ability to hold oil increased from T0 to T8 sample and decreased from T8 to T12 sample. Similarly, water holding capacity increased from T0 to T4 sample and began to gradually decrease from T4 to T12 sample. This trend was similar to study of Jung [24]. This may be due to larger, deeper holes and grooves formed both on the surface and inside starch granules as the hydrolysis time increased. Therefore, the specific surface area also increased when increasing hydrolysis time [20]. However, when hydrolysis reached a certain time, the holes and grooves were merged, reducing the specific surface area of starch granules [25]. Therefore, the ability to hold oil and water was also reduced. Moreover, the structure of starch granules was broken down into fragments when starches were hydrolyzed excessively [2]. This led to holding capacity significantly reduced.



TABLE I. THE OIL HOLDING CAPACITY AND WATER HOLDING CAPACITY OF CORN STARCH AT DIFFERENT HYDROLYSIS TIMES

Sample	OHC (g/g)	WHC (g/g)
T0	1,534 ± 0,035 <sup>a</sup>	1,160 ± 0,035 <sup>a</sup>
T4	1,784 ± 0,028 <sup>bc</sup>	1,949 ± 0,040 <sup>e</sup>
T6	1,836 ± 0,034 <sup>c</sup>	1,769 ± 0,064 <sup>d</sup>
T8	1,925 ± 0,051 <sup>d</sup>	1,566 ± 0,027 <sup>c</sup>
T12	1,744 ± 0,026 <sup>b</sup>	1,368 ± 0,031 <sup>b</sup>

Values are the mean and standard deviation of three samples. Mean values with different letters in the same column are significantly different ( $p \leq 0.05$ ).

### Effect of hydrolysis time on solubility and swelling of starch granules

TABLE II. RESULTS OF SOLUBILITY AND SWELLING POWER OF CORN STARCH AT DIFFERENT HYDROLYSIS TIMES

Sample	SB (%)	SP (g/g)
T0	1,867 ± 0,058 <sup>a</sup>	1,371 ± 0,023 <sup>a</sup>
T4	11,467 ± 0,208 <sup>b</sup>	2,596 ± 0,027 <sup>e</sup>
T6	12,633 ± 0,208 <sup>c</sup>	2,376 ± 0,052 <sup>d</sup>
T8	13,367 ± 0,322 <sup>d</sup>	2,177 ± 0,082 <sup>c</sup>
T12	15,267 ± 0,551 <sup>e</sup>	1,616 ± 0,035 <sup>b</sup>

Values are the mean and standard deviation of three samples. Mean values with different letters in the same column are significantly different ( $p \leq 0.05$ ).

Research results showed that the solubility of hydrolyzed starch was higher than native starch. This was explained by the reduced degree of intermolecular bonds in starch granules during hydrolysis due to glucan chains being cut by enzyme amylase. At the same time, shorter molecules were created and easily dispersed in water [26]. Therefore, the solubility of hydrolyzed starch was increased. Similar results had been reported by Dura [4], Uthumporn [27]. Table 2 also indicated that when hydrolysis time increased, the solubility of the hydrolyzed starch also increased. This may be due to the polymerization degree of amylose and amylopectin chains which decreased with increasing hydrolysis time. Therefore, the molecular size of glucan chains also decreased and the solubility in water increased [28].

Swelling power demonstrated water holding capacity of starch structure and related to amylose content, amylopectin molecular structure and micelle of starch granules [29]. Amylopectin molecule played a major role in starch swelling ability [30]. Besides that, swelling ability was also related to the structural integrity of starch granules. The structural integrity was mainly influenced by the level of interaction between glucan chains (Amylose - Amylose, Amylose - Amylopectin, Amylopectin - Amylopectin) in the crystalline and amorphous region; as well as the arrangement of glucan chains in the crystalline region [31]. The crystalline region in the starch granule was formed by the association of long amylopectin chains. This combination enhanced the structural strength and reduced native starch's swelling [32]. In this study, swelling of hydrolyzed starch was higher than native starch (Table 2). This result was similar to Dura [4] research. However, as the hydrolysis time increased, the starch's swelling capacity was decreased. The reason may be due to the degree of hydrolysis increased when hydrolysis time was long so the structure of amylopectin molecular was shorter. This led to a decrease in swelling capacity of starch granules as Gomand [33] reported.

### Effect of hydrolysis time on gel viscosity

After finishing gelatinization and cooling to 50°C, the viscosity of native starch was higher than hydrolyzed starch. The decrease in gel viscosity after hydrolysis was due to starch molecular was broken down into smaller particle size, thereby reducing in gel viscosity [27]. Results from Table 3 also showed that viscosity between hydrolyzed starch samples was no different. This may be due to big difference in viscosity value between native starch sample and hydrolyzed starch sample. Therefore, to more accurately the change in viscosity between hydrolyzed starch samples, results of changes in viscosity of hydrolyzed starch samples were presented in Table 4.

TABLE III. VISCOSITY OF NATIVE AND HYDROLYZED STARCH

Sample	0h	2h	4h	6h
T0	634,67 ± 19,58 <sup>b</sup>	682,47 ± 21,30 <sup>c</sup>	730,57 ± 27,32 <sup>d</sup>	771,60 ± 23,37 <sup>e</sup>
T4	12,41 ± 0,46 <sup>a</sup>	12,48 ± 0,37 <sup>a</sup>	12,75 ± 0,23 <sup>a</sup>	13,10 ± 0,63 <sup>a</sup>
T6	10,51 ± 0,32 <sup>a</sup>	10,67 ± 0,14 <sup>a</sup>	11,28 ± 0,18 <sup>a</sup>	11,54 ± 0,37 <sup>a</sup>
T8	8,28 ± 0,25 <sup>a</sup>	8,40 ± 0,42 <sup>a</sup>	8,69 ± 0,34 <sup>a</sup>	8,82 ± 0,43 <sup>a</sup>
T12	6,23 ± 0,15 <sup>a</sup>	6,35 ± 0,31 <sup>a</sup>	6,49 ± 0,14 <sup>a</sup>	6,64 ± 0,20 <sup>a</sup>

Values are the mean and standard deviation of three samples. Mean values with different letters in the same column are significantly different ( $p \leq 0.05$ ).

TABLE IV. VISCOSITY OF HYDROLYZED STARCH

Sample	0h	2h	4h	6h
T4	12,41 ± 0,46 <sup>e</sup>	12,48 ± 0,37 <sup>e</sup>	12,75 ± 0,23 <sup>ef</sup>	13,10 ± 0,63 <sup>f</sup>
T6	10,51 ± 0,32 <sup>c</sup>	10,67 ± 0,14 <sup>c</sup>	11,28 ± 0,18 <sup>d</sup>	11,54 ± 0,37 <sup>d</sup>
T8	8,2 ± 0,25 <sup>b</sup>	8,40 ± 0,42 <sup>b</sup>	8,69 ± 0,34 <sup>b</sup>	8,82 ± 0,43 <sup>b</sup>
T12	6,23 ± 0,15 <sup>a</sup>	6,35 ± 0,31 <sup>a</sup>	6,49 ± 0,14 <sup>a</sup>	6,64 ± 0,20 <sup>a</sup>

Values are the mean and standard deviation of three samples. Mean values with different letters in the same column are significantly different ( $p \leq 0.05$ ).

Results showed that, after the end of gelatinization and cooling, viscosity value of hydrolyzed starch decreased with increasing hydrolysis time. This may be due to the structure of glucan chains was cut into smaller particles increased as hydrolysis time increasing [22]. However, after 6 hours of storage, viscosity tended to increase with storage time. This was due to the degeneration process that increased viscosity and formation of starch gel [34].

#### Effect of hydrolysis time on the transmittance of gel

Gel clarity is one of the important factors affecting starch quality [35]. The higher transmittance had the clearer gel [36]. Transmittance was significantly affected by factors such as size, structure and amylose content of starch granules [29, 37]. Low amylose content led to increase transmittance of gel [38]. The change in transmittance of native and hydrolyzed starch after gelatinization and storage for 5 days at 4°C as shown in Fig. 4.

Transmittance of hydrolyzed starch and native starch decreased during 5 days of storage. Increased hydrolysis time lead to a sharp decrease in transmittance value. Transmittance value of T0 was about 7 times higher than T12 sample. The decrease in transmittance during storage due to starch degradation [11, 39] and degree of degradation also increased with storage time [40]. According to Gani [32], aggregation and recrystallization of amylopectin reduced transmittance.

Besides that, reorganization and aggregation of amylose also reduced the transmittance of starch during cold storage [41].

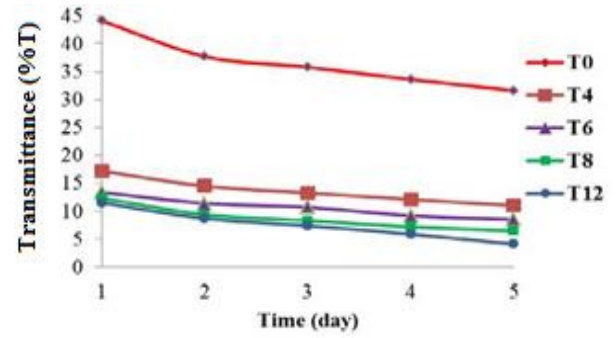


Figure 4. Transmittance of native corn starch and hydrolyzed corn starch

#### Effect of hydrolysis time on CIE L\*a\*b\* of starch

Results in Table 5 indicated that the L\* value of native starch samples was lower than hydrolyzed starch samples. In addition, L\* value also decreased with increasing hydrolysis time. The a\* (red) and WI (whiteness) values of hydrolyzed starch samples were lower than native starch samples. In contrast, b\* (yellow) tended to increase while WI tended to decrease as hydrolysis time increased. The difference in color  $\Delta E$  of the hydrolyzed starch samples was mostly in the range of  $\Delta E < 1$  except T12 sample that fluctuated within  $1 < \Delta E < 2$ . It could be concluded that the difference in color of starch samples after hydrolysis could not be observed visually [42]. Besides that, difference in color tended to increase with increasing hydrolysis time.

TABLE V. CIE L\* A\* B\* COLORIMETRY VALUE OF CORN STARCH AT DIFFERENT HYDROLYSIS TIMES

Sample	L*	a*	b*	WI	$\Delta E$
T0	98,39±0,08 <sup>d</sup>	-0,22±0,00 <sup>c</sup>	2,61±0,05 <sup>a</sup>	96,93±0,00 <sup>e</sup>	
T4	98,38±0,02 <sup>d</sup>	-0,23±0,01 <sup>c</sup>	2,72±0,02 <sup>b</sup>	96,82±0,02 <sup>d</sup>	0,12
T6	98,26±0,04 <sup>c</sup>	-0,26±0,02 <sup>b</sup>	2,98±0,04 <sup>c</sup>	96,53±0,03 <sup>c</sup>	0,40
T8	98,01±0,04 <sup>b</sup>	-0,28±0,01 <sup>b</sup>	3,46±0,01 <sup>d</sup>	96,04±0,02 <sup>b</sup>	0,91
T12	97,51±0,01 <sup>a</sup>	-0,31±0,00 <sup>a</sup>	4,33±0,00 <sup>e</sup>	94,99±0,01 <sup>a</sup>	1,94

Values are the mean and standard deviation of three samples. Mean values with different letters in the same column are significantly different ( $p \leq 0.05$ )

#### IV. CONCLUSION

Hydrolysis by mixture of  $\alpha$ -amylase and amyloglucosidase enzymes changed characteristics and some properties as well as the technological features of corn starch. It broke down the structure of glucan chains into shorter chains, reduced specific gravity of starch molecules and broke the tight structure of starch granules. Therefore, hydrolyzed corn starch samples had higher ability to hold oil, water, solubility, and swelling than native corn starch. The ability to hold oil, water and swelling tended to decrease with increasing hydrolysis time; but the solubility tended to the opposite. Research results also indicated that whiteness as well as viscosity and transmittance of hydrolyzed starch were lower than native starch; and tended to decrease with increasing hydrolysis time.

#### ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

#### REFERENCES

- [1] G. H. P. Te Wierik, J. Bergsma, A. W. Arends-Scholte, T. Boersma, A. C. Eissens, and C. F. Lerk, "A new generation of starch products as excipient in pharmaceutical tablets. I. Preparation and binding properties of high surface area potato starch products". *International journal of pharmaceutics*, vol. 134(1-2), pp. 27-36, 1996.
- [2] B. Zhang, D. Cui, M. Liu, H. Gong, Y. Huang, and F. Han, "Corn porous starch: Preparation, characterization and adsorption property". *International journal of biological macromolecules*, vol. 50(1), pp. 250-256, 2012

- [3] F. Gao, D. Li, C. H. Bi, Z. H. Mao, and B. Adhikari, "Preparation and characterization of starch crosslinked with sodium trimetaphosphate and hydrolyzed by enzymes". *Carbohydrate polymers*, vol. 103, pp. 310-318, 2014.
- [4] A. Dura, W. Błaszczak, and C. M. Rosell, "Functionality of porous starch obtained by amylase or amyloglucosidase treatments". *Carbohydrate polymers*, vol. 101, pp. 837-845, 2014a.
- [5] P. C. Okwuenu, K. U. Agbo, A. L. Ezugwu, S. O. Eze and F. C. Chilaka, "Effect of Divalent Metal Ions on Glucoamylase Activity of Glucoamylase isolated from *Aspergillus niger*". *Ferment Technol*, vol. 6, pp 141-45, 2017
- [6] I. Goñi, A. Garcia-Alonso, and F. Saura-Calixto, "A starch hydrolysis procedure to estimate glycemic index". *Nutrition Research*, vol. 17(3), pp 427-437, 1997.
- [7] C. Sarangapani, R. Thirumdas, Y. Devi, A. Trimukhe, R. R. Deshmukh, and U. S. Annapure, "Effect of low-pressure plasma on physico- chemical and functional properties of parboiled rice flour". *LWT-Food Science and Technology*, vol. 69, pp. 482-489, 2016.
- [8] J. Y Kim, and K. C. Huber, "Corn starch granules with enhanced load-carrying capacity via citric acid treatment". *Carbohydrate polymers*, vol. 91(1), pp. 39-47, 2013.
- [9] A. M. Amini, S. M. A. Razavi, and S. A. Mortazavi, "Morphological, physicochemical, and viscoelastic properties of sonicated corn starch". *Carbohydrate polymers*, vol. 122, pp. 282-292, 2015.
- [10] L. A. Bello-Pérez, K. Meza-León, S. Contreras-Ramos, and O. Paredes-Lopez, "Functional properties of corn, banana and potato starch blends". *Acta científica venezolana*, vol. 52(1), pp. 62-67, 2001.
- [11] I. A. Wani, S. D. Sogi, A. A. Wani, B. S. Gill, and U. S. Shivhare, "Physico-chemical properties of starches from Indian kidney bean (*Phaseolus vulgaris*) cultivars". *International journal of food science & technology*, vol. 45(10), pp. 2176-2185, 2010.
- [12] E. Campechano-Carrera, A. Corona-Cruz, L. Chel-Guerrero, and D. Betancur-Ancona, "Effect of pyrodextrinization on available starch content of Lima bean (*Phaseolus lunatus*) and Cowpea (*Vigna unguiculata*) starches". *Food hydrocolloids*, vol. 21(3), pp. 472-479, 2007.
- [13] L. Atarés, J. Bonilla, and A. Chiralt, "Characterization of sodium caseinate-based edible films incorporated with cinnamon or ginger essential oils". *Journal of Food Engineering*, vol. 100(4), pp. 678-687, 2010.
- [14] C. I. K. Diop, H. L. Li, B. J. Xie, and J. Shi, "Impact of the catalytic activity of iodine on the granule morphology, crystalline structure, thermal properties and water solubility of acetylated corn (*Zea mays*) starch synthesized under microwave assistance". *Industrial crops and products*, vol. 33(2), pp. 302-309, 2011.
- [15] J. E. Fannon, J. M. Shull, and J. N. BeMILLER, "Interior channels of starch granules". *Cereal Chemistry*, vol. 70, pp. 611-611, 1993.
- [16] D. J. Gallant, B. Bouchet, and P. M. Baldwin, "Microscopy of starch: evidence of a new level of granule organization". *Carbohydrate polymers*, vol. 32(3-4), pp. 177-191, 1997.
- [17] K. C. Huber, and J. N. BeMiller, "Channels of maize and sorghum starch granules". *Carbohydrate Polymers*, vol. 41(3), pp. 269-276, 2000.
- [18] W. Helbert, M. Schülein, and B. Henrissat, "Electron microscopic investigation of the diffusion of *Bacillus licheniformis*  $\alpha$ -amylase into corn starch granules". *International Journal of Biological Macromolecules*, vol. 19(3), pp. 165-169, 1996.
- [19] S. Dhital, F. J. Warren, B. Zhang, and M. J. Gidley, "Amylase binding to starch granules under hydrolysing and non-hydrolysing conditions". *Carbohydrate polymers*, vol. 113, pp. 97-107, 2014.
- [20] G. Chen, and B. Zhang, "Hydrolysis of granular corn starch with controlled pore size". *Journal of cereal science*, vol. 56(2), pp. 316-320, 2012
- [21] M. Sujka, and J. Jamroz, "Starch granule porosity and its changes by means of amylolysis". *International agrophysics*, vol. 21(1), pp. 107, 2007
- [22] P. Aggarwal, and D. Dollimore, "A thermal analysis investigation of partially hydrolyzed starch". *Thermochimica Acta*, vol. 319(1-2), pp. 17-25, 1998.
- [23] Y. Benavent-Gil, and C. M. Rosell, "Comparison of porous starches obtained from different enzyme types and levels". *Carbohydrate polymers*, vol. 157, pp. 533-540, 2017.
- [24] Y. S. Jung, B. H. Lee, and S. H. Yoo, "Physical structure and absorption properties of tailor-made porous starch granules produced by selected amylolytic enzymes". *PloS one*, vol. 12(7), e0181372, 2017
- [25] Y. Chen, S. Huang, Z. Tang, X. Chen, and Z. Zhang, "Structural changes of cassava starch granules hydrolyzed by a mixture of  $\alpha$ -amylase and glucoamylase". *Carbohydrate Polymers*, vol. 85(1), pp. 272-275, 2011
- [26] M. Lauro, K. Poutanen, and P. Forsell, "Effect of partial gelatinization and lipid addition on  $\alpha$ -amylolysis of barley starch granules". *Cereal chemistry*, vol. 77(5), pp. 595-601, 2000.
- [27] U. Uthumporn, I. S. Zaidul, and A. A. Karim, "Hydrolysis of granular starch at sub-gelatinization temperature using a mixture of amylolytic enzymes". *Food and Bioproducts Processing*, vol. 88(1), pp. 47-54, 2010.
- [28] T. Tukomane, P. Leerapongnun, S. Shobsngob, and S. Varavinit, "Preparation and characterization of annealed enzymatically hydrolyzed tapioca starch and the utilization in tableting". *Starch-Stärke*, 59(1), 33-45, 2007.
- [29] N. Singh, K. S. Sandhu, and M. Kaur, "Characterization of starches separated from Indian chickpea (*Cicer arietinum* L.) cultivars". *Journal of Food Engineering*, vol. 63(4), pp. 441-449, 2004.
- [30] R. F. Tester, and W. R. Morrison, "Swelling and gelatinization of cereal starches. I. Effects of amylopectin, amylose, and lipids". *Cereal chem*, vol. 67(6), pp. 551-557, 1990.
- [31] V. Vamadevan, E. Bertoft, and K. Seetharaman, "On the importance of organization of glucan chains on thermal properties of starch". *Carbohydrate polymers*, vol. 92(2), pp. 1653-1659, 2013.
- [32] A. Gani, S. M. Wani, F. A. Masoodi, and R. Salim, "Characterization of rice starches extracted from Indian cultivars". *Food Science and Technology International*, vol. 19(2), pp. 143-152, 2013.
- [33] S. V. Gomand, L. Lamberts, R. G. F. Visser, and J. A. Delcour, "Physicochemical properties of potato and cassava starches and their mutants in relation to their structural properties". *Food Hydrocolloids*, vol. 24(4), pp. 424-433, 2010.
- [34] S. Wang, C. Li, L. Copeland, Q. Niu, and S. Wang, "Starch retrogradation: A comprehensive review". *Comprehensive Reviews in Food Science and Food Safety*, vol. 14(5), pp. 568-585, 2015.
- [35] S. E. Alemán, A. O. Ramírez, E. E. Manzanilla, R. E. Guzmán, and E. E. Pérez, "Functional and nutritional characterization of native and modified starches from bananas hybrids". *Starch-Stärke*, vol. 67(5-6), pp. 459-469, 2015.
- [36] B. A. Ashwar, A. Shah, A. Gani, S. A. Rather, S. M. Wani, I. A. Wani, and A. Gani, "Effect of gamma irradiation on the physicochemical properties of alkali-extracted rice starch". *Radiation Physics and Chemistry*, vol. 99, pp. 37-44, 2014.
- [37] S. Yu, Y. Ma, L. Menager, and D. W. Sun, "Physicochemical properties of starch and flour from different rice cultivars". *Food and bioprocess technology*, vol. 5(2), pp. 626-637, 2012
- [38] J. M. Swinkels, *Sources of starch, its chemistry and physics. In: Starch Conversion Technology*. New York: Marcel Dekker, 1985, pp. 15-46.
- [39] C. K. Reddy, M. Suriya, P. V. Vidya, K. Vijina, and S. Haripriya, "Effect of  $\gamma$ -irradiation on structure and physico-chemical properties of *Amorphophallus paeoniifolius* starch". *International journal of biological macromolecules*, vol. 79, pp. 309-315, 2015.
- [40] A. Gani, M. Bashir, S. M. Wani, and F. A. Masoodi, "Modification of bean starch by  $\gamma$ -irradiation: Effect on functional and morphological properties". *LWT-Food Science and Technology*, vol. 49(1), pp. 162-169, 2012.
- [41] T. F. Achille, A. Nrsquo, G. Georges and K. Alphonse, "Contribution to light transmittance modelling in starch media". *African Journal of Biotechnology*, vol. 6(5), pp. 569-575, 2007.
- [42] W. S., Mokrzycki and M. Tatol, "Colour difference  $\Delta E$ -A survey". *Machine Graphics and Vision*, vol. 20(4), pp. 383-411, 2011.

# PyPSA-VN: An open model of the Vietnamese electricity system

Markus Schlott  
Frankfurt Institute  
for Advanced Studies  
Frankfurt, Germany  
schlott@fias.uni-frankfurt.de

Bruno Schyska  
DLR Institute  
of Networked Energy Systems  
Oldenburg, Germany  
bruno.schyska@dlr.de

Dinh Thanh Viet  
University  
of Da Nang  
Da Nang, Vietnam  
dtviet@ac.udn.vn

Vo Van Phuong  
University  
of Da Nang  
Da Nang, Vietnam  
vanphuong0812@gmail.com

Duong Minh Quan  
University  
of Da Nang  
Da Nang, Vietnam  
dmquanbk03@gmail.com

Ma Phuoc Khanh  
Central Region  
Load Dispatch Centre  
Da Nang, Vietnam  
khanhmp.a3@nldc.evn.vn

Fabian Hofmann  
Frankfurt Institute  
for Advanced Studies  
Frankfurt, Germany  
hofmann@fias.uni-frankfurt.de

Lueder von Bremen  
DLR Institute  
of Networked Energy Systems  
Oldenburg, Germany  
lueder.von.bremen@dlr.de

Detlev Heinemann  
University  
of Oldenburg  
Oldenburg, Germany  
detlev.heinemann@uol.de

Alexander Kies  
Frankfurt Institute  
for Advanced Studies  
Frankfurt, Germany  
kies@fias.uni-frankfurt.de

**Abstract**—Vietnam has tremendous potential for renewable energy. However, its renewable generation mix remains in its infancy, while demand grows rapidly year after year. In this work, we present an open dataset to model the Vietnamese electricity system implemented in Python for Power System Analysis (PyPSA) and use it to analyse a Vietnamese power development plan until 2030. We analyse two different meteorological datasets, the ERA5 and MERRA-2 reanalyses and use them to cost-optimize a future Vietnamese power system under an official power development plan as boundary condition and compare the findings. We show that differences in the results are small for power systems mostly based on conventional thermal generation.

**Index Terms**—Vietnamese Power System, Energy System Analysis, Open Data, PyPSA, ERA5, MERRA-2

## I. INTRODUCTION

Despite having considerable unexplored potential for diverse renewable energy resources, Vietnam plans to power its strong economic growth mainly based on fossil power sources. This is not only true for Vietnam, but also for other countries in Southeast Asia [1]. The foreseeable strong growth in electricity demand as well as the unique geographical location and features of Vietnam make a careful planning of the power system expansion necessary, because weak grid capacity is a major barrier for investments into renewable generation in Vietnam [2].

In this work, we present an open dataset of the Vietnamese electricity system and use it to study the cost-optimal ex-

pansion of the Vietnamese power system with respect to an official power development plan (PDP), which is mainly based on an expansion of conventional energy carriers. In the PDP scenario, the renewable power share of installed capacities grows in the two 5-year periods from 2020 to 2030 from around 3% to 15%.

To model generation from the renewable sources of wind and solar photovoltaics (PV), an appropriate understanding of underlying weather situations is mandatory. Country-wide renewable power system studies usually rely on weather data from reanalyses [3] and since renewable energy sources depend on the weather and are therefore not dispatchable, their integration is a challenging task. Among proposed solutions to integrate renewables into power systems are: using the complementarity of different renewables [4], [5] such as wind and PV [6], [7], wind/PV/hydro [8], wind/hydro [9], PV/hydro [10], storage [11]–[13], transmission grid extensions [14]–[16], demand-side management [17]–[19], system-friendly renewables [20]–[22] or sector-coupling [23], [24].

The potential for renewables and a renewable power system in Vietnam have already been studied in a number of recent works: Polo et al. [25] studied solar potentials using satellite and GIS-based information, Nguyen et al. [26]–[28] studied the optimisation of a renewable Vietnamese power system with a special emphasis on wind power. Kies et al. [29] studied the large-scale integration of renewables in a future power system in a simplified setting. Huber et al. [30] did

not only investigate Vietnam, but instead embedded it within the ASEAN region. Do and Hoffmann [31] studied three scenarios for the Vietnamese power system expansion planning from 2018 to 2030 and found that coal-fired installations will likely contribute a significantly reduced share to the overall generation mix more than envisioned.

We compare the results using two different weather datasets: ERA5 [32] is a climate reanalysis dataset developed by the Copernicus Climate Change Service. MERRA-2 [33] is a reanalysis dataset provided by NASA. Both datasets have been extensively used in several power system-related studies. Olason [34] compared ERA5 with MERRA-2 for wind power modelling and found ERA5 to perform considerably better for country-wide output as well as for individual wind turbines. Aniskevich et al. [35] used ERA5 to investigate wind energy resources in Latvia and Schindler et al. [36] used ERA5 to estimate wind energy yields in Germany.

In this paper we present an open dataset for the electricity system of Vietnam. The model is available on github (<https://github.com/fiasresna/pypsa-vn/>). In the next sections, we describe the cost-optimisation model and the data used and created and finally, we present the results as well as a summary.

## II. METHODOLOGY

The topology of the network model is based upon the Vietnamese high voltage transmission grid as well as on its power plant fleet as it exists today. Data on the latter were taken from the Global Power Plant Database [37]. One optimisation run covers the period of a full year, where the weather datasets for Vietnam are taken from ERA5 and MERRA-2 for the year 2015. All results in this work were obtained using the software-toolbox Python for Power System Analysis [38].

### A. Objective and Constraints

The optimisation objective is to find the cost-optimal network solution for the mentioned power system under various constraints. The problem itself is a linear minimisation of total system cost [38] and reads:

$$\min_{g,G,F} \left( \sum_{n,s} c_{n,s} \cdot G_{n,s} + \sum_l c_l \cdot F_l + \sum_{n,s,t} o_s \cdot g_{n,s,t} \right) \quad (1)$$

The objective consists of capital costs  $c_{n,s}G_{n,s}$  for installed capacity of a carrier  $s$  at node  $n$ , capital costs  $c_l F_l$  for transmission capacity at line  $l$  and marginal costs of generation  $o_s g_{n,s,t}$  for energy generation of carrier  $s$  at node  $n$  and time  $t$ .

The cost assumptions for all relevant technologies are based on own assumptions, they are annualised with a discount rate of 7% and given in Table 1. Capital cost assumptions are a critical aspect of power system expansion models and can significantly effect results [39]. Reducing financing risks in the renewable energy sector via political means could reduce the capital need for renewable investment [40].

The optimisation problem is subject to various constraints; most important is the nodal power balance among generation,

demand and flow, as well as dispatch constraints with respect to generation:

$$\begin{aligned} \sum_s g_{n,s}(t) - d_n(t) &= \sum_l K_{n,l} \cdot f_l(t) \quad \forall \quad n, t \\ g_{n,s}^-(t) \cdot G_{n,s} &\leq g_{n,s}(t) \leq g_{n,s}^+(t) \cdot G_{n,s} \quad \forall \quad n, t \end{aligned} \quad (2)$$

In the equation for nodal power balance,  $d_n(t)$  is the demand at node  $n$  and time  $t$ ,  $K_{n,l}$  the networks incidence matrix and  $f_l(t)$  the actual power flow on line  $l$  and time  $t$ . The energy generation in the dispatch constraint is limited downwards by its minimal generation potential  $g_{n,s}^-(t)$  and limited upwards by its maximum generation potential  $g_{n,s}^+(t)$ , where  $G_{n,s}$  describes the maximum installable capacity of a carrier type  $s$  at node  $n$ . Both constraints also apply for storage units.

Storage units have to obey additional constraints given by the state of charge (soc) equations, which describe the charging and discharging behaviour:

$$\begin{aligned} \text{soc}_{n,s}(t) &= \eta_0 \cdot \text{soc}_{n,s}(t-1) + \eta_1 \cdot g_{n,s,\text{store}}(t) \\ &\quad - \eta_2^{-1} \cdot g_{n,s,\text{dispatch}}(t) + \text{inflow}_{n,s}(t) \\ &\quad - \text{spillage}_{n,s}(t) \quad \forall \quad n, s, t > 1 \\ \text{soc}_{n,s}(t_0) &= \text{soc}_{n,s}(t|t) \quad \forall \quad n, s, t \end{aligned} \quad (3)$$

Charging efficiencies are described by  $\eta$ , where  $\eta_0$  describes standing losses,  $\eta_1$  efficiencies of storage uptake and  $\eta_2$  efficiencies of storage dispatch. The term inflow covers any external inflow, e.g. water runoff into hydro reservoirs and spillage any spillage.

Total generation capacities  $G_{n,s}$  for a carrier type  $s$  at node  $n$  as well as total line capacities  $F_l$  for a line  $l$  are restricted to a certain interval

$$\begin{aligned} G_{n,s}^{\min} &\leq G_{n,s} \leq G_{n,s}^{\max} \\ F_l^{\min} &\leq F_l, \end{aligned} \quad (4)$$

where the superscript min denotes the minimum capacity and max the maximum capacity. For onshore wind a maximum capacity of 10 MW/km<sup>2</sup> and for solar PV of 170 MW/km<sup>2</sup> is assumed. Existing hydro and run-of-river plants are allowed to be expanded by 50% of today's capacity value. The expansion however is treated as run-of-river plant types solely without any further assumptions on the general dam potentials throughout the Vietnamese landscape. The minimum line capacity represents Vietnam's high voltage power grid as it exists today; an unrestricted expansion potential is considered.

For the flow  $f_l$  on a line  $l$  featuring a series reactance  $x_l$ , the linearised power flow equations with respect to the voltage angles among the interconnected buses  $i$  and  $j$  have to be obeyed:

$$\begin{aligned} f_l &= \frac{\theta_i - \theta_j}{x_l} \\ |f_l(t)| &\leq F_l \quad \forall \quad l \end{aligned} \quad (5)$$

In addition, the absolute power flow  $|f_l(t)|$  on a line cannot exceed its nominal line capacity  $F_l$ .



TABLE I: Cost assumptions used for the simulation, based on own assumptions.

Technology	Capital Cost [Euro/MW/a]	Marginal Cost [Euro/MWh]	Efficiency Dispatch
Hard Coal	222,000	30.0	
Oil	255,300	91.6	
OCGT	63,800	76.80	
Onshore Wind	140,400	0.015	
Solar PV	65,920	0.01	
Run-Of-River	211,200	5.00	
Hydro Reservoir	211,200	5.00	0.9
Bioenergy	256,000	72.90	
Transmission Lines	16,450	Euro/MWkm/a	

### III. RENEWABLE GENERATION DATA

To model renewable generation time-series, local weather data on wind speeds, irradiation, temperature and water runoff are taken from the global ERA5 as well as the MERRA-2 reanalysis for the reference year 2015. ERA5 is a climate reanalysis dataset covering the period 1950 to present, which is being provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). It has a spatial resolution of approximately 31 km and an hourly temporal resolution for most meteorological variables. Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2) is developed and maintained by NASA. It has a spatial resolution of approximately 50 km and hourly temporal resolution. The meteorological data from both reanalyses is provided on a spatial grid and, after conversion to power, the generation time-series are aggregated to the corresponding nodes of the network using the principle of Voronoi cells (nearest neighbour). For MERRA-2, we have not processed the data by ourselves, but instead retrieved the relevant renewable generation time-series pointwise from the renewables.ninja [41] project. While using the same technical specifications (turbine type, panel type, angle/tilt, etc.), the discrepancy between point-like values for MERRA-2 compared to aggregated grid values in the case of ERA5 should be kept in mind. Renewable generation time-series are calculated for the energy carrier types onshore wind, solar PV, and hydro; the specific methodology applied to these carrier types is described in the following.

#### A. Onshore Wind

For both datasets, wind speed is directly provided; in the case of ERA5 at a height level of 100 m, and in the case of MERRA-2 at a height level of 50 m. As a first step, wind speeds are converted to 80 m using a wind power log profile:

$$\frac{u(z_2)}{u(z_1)} = \frac{\log(\frac{z_2}{z_0})}{\log(\frac{z_1}{z_0})}, \quad (6)$$

where  $u$  is the wind speed,  $z_2$  is the chosen hub height (80 m) and  $z_1$  is the height at which wind speeds are given (100 m / 50 m).  $z_0$  is the surface roughness length, which is provided as a static quantity. Wind speeds at hub height are then converted to power using the power curve of a Vestas V112 3 MW turbine for onshore locations (the offshore wind potential of

Vietnam is not incorporated in this work). To better match the actual wind feed-in data the power curve is additionally being smoothed with a Gaussian kernel [42] in the case of ERA5.

#### B. Solar PV

For PV power two different approaches are used. In the case of MERRA-2 the generation is determined by the solar PV model from renewables.ninja, which incorporates system losses (10%), Tilt (35°) and Azimuth (180°, facing south). In the case of ERA5 the solar PV model from the Python package atlite is invoked; the latter's methodology shall be shortly described here: First, the downwelling shortwave irradiation is split into direct and diffuse horizontal irradiation using the Reindl Clearsky model [43]. Both parts are then rotated on the tilted solar panel surface. The tilted rotated irradiation consists of three parts. The first one is the tilted direct irradiation (TDI), which simply describes the geometrically rotated direct horizontal irradiation. The diffuse tilted irradiation (DTI) is the rotated diffuse irradiation, where the rotation is executed subsequent to the Hay-Davies model [44]. The ground tilted irradiation (GTI) is determined from the ground albedo invoking the upwelling shortwave irradiation. The total irradiation on the tilted plane (TTI) is finally given as the sum of its components:

$$TTI = TDI + DTI + GTI. \quad (7)$$

Potential power generation from PV is then calculated using an effective solar panel model from Huld et al. [45] for the CSi panel, where AC power per installed capacity unit is given as a function of the normalised total tilted irradiation:

$$\bar{g}^{PV} = TTI_N \cdot \eta \cdot \eta_{inv} \quad (8)$$

The efficiency  $\eta$  is given by

$$\begin{aligned} \eta = 1 &+ k_1 \cdot \ln(TTI_N) + k_2 \cdot \ln(TTI_N)^2 \\ &+ T_N \cdot \left( k_3 + k_4 \cdot \ln(TTI_N) + k_5 \cdot \ln(TTI_N)^2 \right) \\ &+ k_6 \cdot T_N^2, \end{aligned} \quad (9)$$

and the normalised total tilted irradiation is given by the panels reference irradiance  $I_R = 1000$  W via

$$TTI_N = \frac{TTI}{I_R}. \quad (10)$$

$T_N$  describes the normalised module temperature as the difference between irradiation-corrected air temperature and the reference standard test temperature  $T_R = 298$  K, given by

$$T_N = (c_1 \cdot T_{\text{amb}} + c_2 \cdot \text{TTI}) - T_R, \quad (11)$$

where the coefficients for the chosen panel are  $c_1 = 1$  and  $c_2 = 0.035 \text{ Km}^2/\text{W}$ .

The quantities  $k_1$  to  $k_6$  are pure device-specific parameters and given by  $k_1 = -0.017162$ ,  $k_2 = -0.040289$ ,  $k_3 = -0.004681/\text{K}$ ,  $k_4 = 0.000148/\text{K}$ ,  $k_5 = 0.000169/\text{K}$  and  $k_6 = 0.000005/\text{K}^2$ . The inverter efficiency  $\eta_{\text{inv}}$  is chosen to be 0.9.

### C. Hydro Power

To model hydro power, we use a potential energy approach based on water runoff data from the ERA5 reanalysis for all runs (including MERRA-2) as described by Kies et al. [46]. Because of inadequate water runoff data in the case of MERRA-2 in Vietnam, we decide to abandon the mentioned variable and instead only invoke the ERA5 one; with respect to this decision we try to focus on effects concerning onshore wind and solar PV generation without fully removing generation from hydro reservoirs as well as run-of-river plants from the network.

Power from hydro sources is derived from the potential gravitational energy of a mass  $m$  relative to the sea level, given by

$$U = m \cdot g \cdot h, \quad (12)$$

where  $g = 9.81 \text{ m/s}^2$  is the gravitational acceleration on Earth and  $h$  the height above sea level. For each grid cell spanning an area  $dA$  in a given country, the inflow into hydro reservoirs and run-of-river plants is calculated as a linear function of the potential energy of the water runoff variable:

$$\text{inflow}(t) = f \cdot g \int_A m(x, y, t) h(x, y) dA. \quad (13)$$

$f$  is a normalization constant that ensures  $\langle I_n^H(t) \rangle = \langle G_n^H(t) \rangle$ , where the latter is today's average hourly generation from hydro plants in the corresponding country. Normalisation data for Vietnam is taken from the international hydropower association [47].

## IV. RESULTS

We study the official Vietnamese Power Development Plan (PDP) evolving from 2020 across 2025 to the year 2030. It takes the expectation on generation capacity increase with respect to several energy carriers but also the expected total load increase into account; all relevant values are given in Tab. II. However, a comparison with real values of solar PV generation (also shown in Tab. II) reveals that installed solar PV capacities in 2019 already surpassed expectations for 2020 by far. To give the model some flexibility, we allow for a variation of  $\pm 10\%$  per carrier type:

$$\sum_n G_{n,s} = G_s^{\text{PDP}} \pm 10\% \quad (14)$$

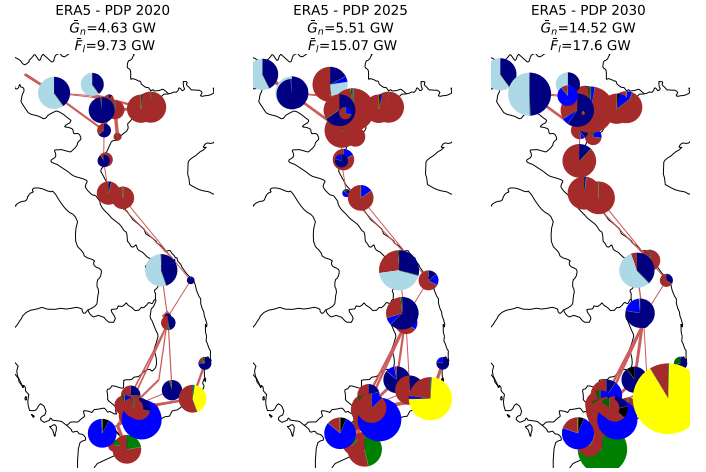


Fig. 1: Cost-optimised distribution of generation facilities in accordance with the Vietnamese PDP for 2020, 2025 and 2030 for ERA5 data. Technologies shown are solar PV (yellow), wind (green), hydro (navy blue), hard coal (brown), OCGT (blue), bio (dark green), oil (black) and run-of-river (light blue). Overall capacities are given in the Table II.  $\bar{G}_n$  refers to the maximum generation capacity of the largest node and  $\bar{F}_l$  to the maximum line capacity of the largest line.

We study the cost-optimal power system using the meteorological datasets from MERRA-2 and ERA5 as input. Fig. 1 and 2 show installed capacities. Overall capacities are given in Table II. The figures reveal that for the PDP scenarios the spatial distributions of capacities look very similar. Solar PV is, in both cases, only installed in the South of Vietnam. The same holds for wind, which can also be found in the South. In the North, a larger share of power is provided by hydro sources.

Taking a look at the dispatch of different technologies for the investigated scenarios, Fig. 3 shows that the situation is dominated by conventional generation. However, dispatched wind energy differs drastically between both scenarios, while solar PV dispatch is similar. For ERA5, dispatch of wind energy is much smaller than in the case of MERRA-2. This indicates a significant effect arising from the input weather data. The missing wind generation is mostly replaced by generation from hard coal.

## V. SUMMARY AND CONCLUSION

In this work, we have presented an open model of the Vietnamese electricity system. The model is available via github and provides data necessary to perform power system studies for a future highly renewable Vietnamese power system.

We have used this data to study the Vietnamese power development plan until 2030 and have shown that differences arising from using different weather datasets, namely the ERA5 and MERRA-2 reanalysis, exist. These differences have a significant effect on the cost-optimised power system design. It would be worth investigating how these differences evolve in a scenario without large shares of conventional generation

TABLE II: Scenario constraints from the Vietnamese Power Development Plan (PDP, published in 2016). Generation capacities are given in MW. Total load increase is given as a factor, which, applied to the initial total load, releases the assumed total load in the specified year. The last column contains installed capacities as of 2019 according to IRENA [48].

Year	2020	2025	2030	2019
Technology [MW]				
Hard Coal	26000	47600	55300	
Oil	1065	1775	2248	
Gas	9000	15000	19000	
Onshore Wind	800	2000	6000	275 (+ 99 offshore)
Solar PV	850	4000	12000	5695
Run-of-River	4850	7850	11050	
Hydro Reservoir	16750	16750	16750	18069 (hydro in total)
Bioenergy	306	559	1407	365
Average Load	1.787	2.722	3.965	

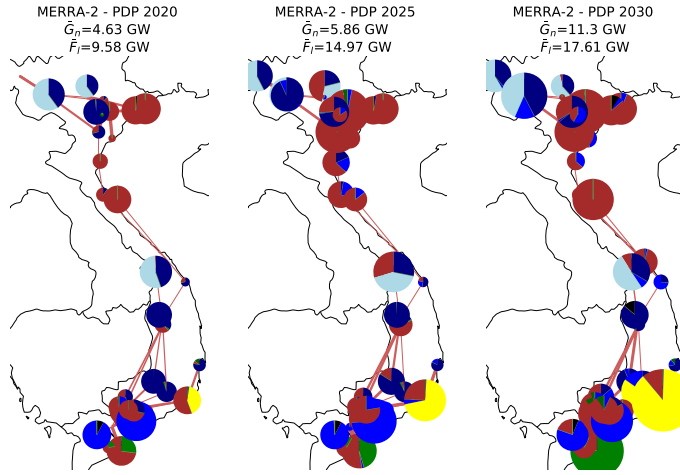


Fig. 2: Cost-optimised distribution of generation facilities in accordance with the Vietnamese PDP for 2020, 2025 and 2030 for MERRA-2 data. Technologies shown are solar PV (yellow), wind (green), hydro (navy blue), hard coal (brown), OCGT (blue), bio (dark green), oil (black) and run-of-river (light blue). Overall capacities are given in table II.  $\bar{G}_n$  refers to the maximum generation capacity of the largest node and  $\bar{F}_l$  to the maximum line capacity of the largest line.

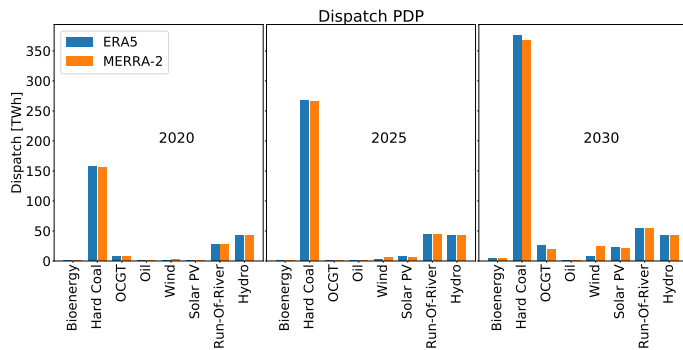


Fig. 3: Aggregated dispatch per technology for different scenarios using ERA5 and MERRA-2 as input.

capacities, i.e. under ambitious CO<sub>2</sub> emission reduction constraints.

#### ACKNOWLEDGEMENTS

This work is part of the R&D Project Analysis of the Large Scale Integration of Renewable Power into the Future Vietnamese Power System financed by Gesellschaft fuer Internationale Zusammenarbeit GmbH (GIZ GmbH).

#### REFERENCES

- [1] R. Mamat, M. Sani, K. Sudhakar *et al.*, “Renewable energy in southeast asia: Policies and recommendations,” *Science of the total environment*, vol. 670, pp. 1095–1102, 2019.
- [2] A. URAKAMI, “Are the barriers to private solar/wind investment in vietnam mainly those that limit network capacity expansion?” 2020.
- [3] J. Jurasz, F. Canales, A. Kies, M. Guezgouz, and A. Beluco, “A review on the complementarity of renewable energy sources: Concept, metrics, application and future research directions,” *Solar Energy*, vol. 195, pp. 703–724, 2020.
- [4] J. Jurasz, A. Beluco, and F. A. Canales, “The impact of complementarity on power supply reliability of small scale hybrid energy systems,” *Energy*, vol. 161, pp. 737–743, 2018.
- [5] N. S. Thomaidis, F. J. Santos-Alamillos, D. Pozo-Vázquez, and J. Usaola-García, “Optimal management of wind and solar energy resources,” *Computers & Operations Research*, vol. 66, pp. 284–291, 2016.
- [6] D. Heide, L. Von Bremen, M. Greiner, C. Hoffmann, M. Speckmann, and S. Bofinger, “Seasonal optimal mix of wind and solar power in a future, highly renewable europe,” *Renewable Energy*, vol. 35, no. 11, pp. 2483–2489, 2010.
- [7] F. Santos-Alamillos, D. Pozo-Vázquez, J. Ruiz-Arias, L. Von Bremen, and J. Tovar-Pescador, “Combining wind farms with concentrating solar plants to provide stable renewable power,” *Renewable Energy*, vol. 76, pp. 539–550, 2015.
- [8] J. Jurasz, P. B. Dabek, B. Kaźmierczak, A. Kies, and M. Wdowikowski, “Large scale complementary solar and wind energy sources coupled with pumped-storage hydroelectricity for lower silesia (poland),” *Energy*, vol. 161, pp. 183–192, 2018.
- [9] F. A. Canales, A. Beluco, and C. A. B. Mendes, “A comparative study of a wind hydro hybrid system with water storage capacity: Conventional reservoir or pumped storage plant?” *Journal of Energy Storage*, vol. 4, pp. 96–105, 2015.
- [10] B. Ming, P. Liu, S. Guo, X. Zhang, M. Feng, and X. Wang, “Optimizing utility-scale photovoltaic power generation for integration into a hydropower reservoir by incorporating long-and short-term operational decisions,” *Applied Energy*, vol. 204, pp. 432–445, 2017.
- [11] S. Weitemeyer, D. Kleinhans, T. Vogt, and C. Agert, “Integration of renewable energy sources in future power systems: The role of storage,” *Renewable Energy*, vol. 75, pp. 14–20, 2015.

- [12] S. Weitemeyer, D. Kleinhans, L. Wienholt, T. Vogt, and C. Agert, "A european perspective: potential of grid and storage for balancing renewable power systems," *Energy Technology*, vol. 4, no. 1, pp. 114–122, 2016.
- [13] D. Heide, M. Greiner, L. Von Bremen, and C. Hoffmann, "Reduced storage and balancing needs in a fully renewable european power system with excess wind and solar power generation," *Renewable Energy*, vol. 36, no. 9, pp. 2515–2523, 2011.
- [14] R. A. Rodriguez, S. Becker, G. B. Andresen, D. Heide, and M. Greiner, "Transmission needs across a fully renewable european power system," *Renewable Energy*, vol. 63, pp. 467–476, 2014.
- [15] F. Steinke, P. Wolfrum, and C. Hoffmann, "Grid vs. storage in a 100% renewable europe," *Renewable Energy*, vol. 50, pp. 826–832, 2013.
- [16] K.-K. Cao, J. Metzendorf, and S. Birbalta, "Incorporating power transmission bottlenecks into aggregated energy system models," *Sustainability*, vol. 10, no. 6, p. e1916, 2018.
- [17] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE transactions on industrial informatics*, vol. 7, no. 3, pp. 381–388, 2011.
- [18] A. Zerrahn and W.-P. Schill, "On the representation of demand-side management in power system models," *Energy*, vol. 84, pp. 840–845, 2015.
- [19] A. Kies, B. U. Schyska, and L. von Bremen, "The demand side management potential to balance a highly renewable european power system," *Energies*, vol. 9, no. 11, p. 955, 2016.
- [20] L. Hirth and S. Müller, "System-friendly wind power: How advanced wind turbine design can increase the economic value of electricity generated through wind power," *Energy Economics*, vol. 56, pp. 51–63, 2016.
- [21] P. Tafarte, A. Kanngießer, M. Dotzauer, B. Meyer, A. Grevé, and M. Millinger, "Interaction of electrical energy storage, flexible bioenergy plants and system-friendly renewables in wind-or solar pv-dominated regions," *Energies*, vol. 13, no. 5, p. 1133, 2020.
- [22] K. Chattopadhyay, A. Kies, E. Lorenz, L. von Bremen, and D. Heinemann, "The impact of different pv module configurations on storage and additional balancing needs for a fully renewable european power system," *Renewable Energy*, vol. 113, pp. 176–189, 2017.
- [23] T. Brown, D. Schlachtberger, A. Kies, S. Schramm, and M. Greiner, "Synergies of sector coupling and transmission reinforcement in a cost-optimised, highly renewable european energy system," *Energy*, vol. 160, pp. 720–739, 2018.
- [24] K. Schaber, F. Steinke, and T. Hamacher, "Managing temporary oversupply from renewables efficiently: Electricity storage versus energy sector coupling in germany," in *International Energy Workshop, Paris*, 2013.
- [25] J. Polo, A. Bernardos, A. Navarro, C. Fernandez-Peruchena, L. Ramirez, M. V. Guisado, and S. Martinez, "Solar resources and power potential mapping in vietnam using satellite-derived and gis-based information," *Energy conversion and management*, vol. 98, pp. 348–358, 2015.
- [26] K. Q. Nguyen, "Impacts of wind power generation and co2 emission constraints on the future choice of fuels and technologies in the power sector of vietnam," *Energy Policy*, vol. 35, no. 4, pp. 2305–2312, 2007.
- [27] Q. K. Nguyen, "Long term optimization of energy supply and demand in vietnam with special reference to the potential of renewable energy," Ph.D. dissertation, Universität Oldenburg, 2005.
- [28] K. Q. Nguyen, "Wind energy in vietnam: Resource assessment, development status and future implications," *Energy Policy*, vol. 35, no. 2, pp. 1405–1413, 2007.
- [29] A. Kies, B. Schyska, D. T. Viet, L. von Bremen, D. Heinemann, and S. Schramm, "Large-scale integration of renewable power sources into the vietnamese power system," *Energy Procedia*, vol. 125, pp. 207–213, 2017.
- [30] M. Huber, A. Roger, and T. Hamacher, "Optimizing long-term investments for a sustainable development of the asean power system," *Energy*, vol. 88, pp. 180–193, 2015.
- [31] T. H. Do and C. Hoffmann, "A power development planning for vietnam under the co2 emission reduction targets," *Energy Reports*, vol. 6, pp. 19–24, 2020.
- [32] K. Hennermann and P. Berrisford, "Era5 data documentation," 2017.
- [33] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim *et al.*, "Merra: Nasa's modern-era retrospective analysis for research and applications," *Journal of climate*, vol. 24, no. 14, pp. 3624–3648, 2011.
- [34] J. Olauson, "Era5: The new champion of wind power modelling?" *Renewable energy*, vol. 126, pp. 322–331, 2018.
- [35] S. Aniskevich, V. Bezrukovs, U. Zandovskis, and D. Bezrukovs, "Modelling the spatial distribution of wind energy resources in latvia," *Latvian Journal of Physics and Technical Sciences*, vol. 54, no. 6, pp. 10–20, 2017.
- [36] D. Schindler and C. Jung, "Copula-based estimation of directional wind energy yield: A case study from germany," *Energy Conversion and Management*, vol. 169, pp. 359–370, 2018.
- [37] L. Byers, J. Friedrich, R. Hennig, A. Kressig, X. Li, C. McCormick, and L. M. Valeri, "A global database of power plants," *World Resour. Inst.*, vol. 18, 2018.
- [38] T. Brown, J. Hörsch, and D. Schlachtberger, "Pypsa: Python for power system analysis," *arXiv preprint arXiv:1707.09913*, 2017.
- [39] B. U. Schyska and A. Kies, "How regional differences in cost of capital influence the optimal design of power systems," *Applied Energy*, vol. 262, p. 114523, 2020.
- [40] C. Klessmann, M. Rathmann, D. de Jager, A. Gazzo, G. Resch, S. Busch, and M. Ragwitz, "Policy options for reducing the costs of reaching the european renewables target," *Renewable Energy*, vol. 57, pp. 390–403, 2013.
- [41] S. Pfenninger and I. Staffell, "Renewables. ninja," URL <https://www.renewables.ninja>, 2016.
- [42] G. B. Andresen, A. A. Søndergaard, and M. Greiner, "Validation of Danish wind time series from a new global renewable energy atlas for energy system analysis," *Energy*, vol. 93, pp. 1074–1088, 2015.
- [43] D. Reindl, W. Beckman, and J. Duffie, "Evaluation of hourly tilted surface radiation models," *Solar energy*, vol. 45, no. 1, pp. 9–17, 1990.
- [44] J. Davies and J. Hay, "Calculation of the solar radiation incident on an inclined surface," in *Proc. First Canadian Solar Radiation Data Workshop*, pp. 59–72.
- [45] T. Huld, G. Friesen, A. Skoczek, R. P. Kenny, T. Sample, M. Field, and E. D. Dunlop, "A power-rating model for crystalline silicon pv modules," *Solar Energy Materials and Solar Cells*, vol. 95, no. 12, pp. 3359–3369, 2011.
- [46] A. Kies, K. Chattopadhyay, L. von Bremen, E. Lorenz, and D. Heinemann, "Restore 2050: Simulation of renewable feed-in for power system studies," Tech. Rep, Tech. Rep., 2016.
- [47] "Hydropower country profile: Vietnam," <https://www.hydropower.org/country-profiles/vietnam>.
- [48] IRENA, "Renewable energy statistics," 2019.

# Customer satisfaction with service quality: An empirical study of An Giang Power Company

Hong-Xuyen Ho Thi

*Faculty of Economics,*

*Ho Chi Minh City University of Technology and Education,*

*Ho Chi Minh City 71307, Vietnam*

*xuyenhth@hcmute.edu.vn*

Lan Anh Nguyen Thi

*Faculty for High Quality Training,*

*Ho Chi Minh City University of Technology and Education,*

*Ho Chi Minh City 71307, Vietnam*

*lananhnt@hcmute.edu.vn*

**Abstract—** The electricity industry is one of the service sectors that plays an important role in Vietnam's economy. Ensuring sustainable power supply for production, business and life is a challenging issue. Thus, this study investigates factors that affect customer satisfaction with quality service about power supply in An Giang Power Company. The result provides the useful information to help managers of An Giang Power Company have an overview of power service situation in order to improve the quality, services and find the solutions to ensure sustainable development for An Giang power company as well as Vietnam's electricity industry.

**Keywords—**customer satisfaction, service quality, power supply service

## I. INTRODUCTION

Nowadays, Vietnam's economy is assessed to achieve a high growth rate in the world. Service industries play an increasingly important role in Vietnam's economy, especially the electricity industry. In a global competitive environment, providing quality services to meet the increasing needs of customers is an important strategy for the survival and success of many enterprises.

Electricity is a special commodity that is produced, transferred and consumed at the same time. The electricity industry is a monopoly industry in Vietnam. As a single unit to provide electricity for the citizen and socio-economic organizations, electricity market has no competition. Since joining the World Trade Organization with the request to improve product quality, service and improve the efficiency of production and business. Therefore, the power supply of the electricity industry must also meet the requirements for improving service quality. Ensuring power supply for production and daily life, especially ensuring sufficient energy supply for economic development, agriculture and services are an urgent issue for the Government of Vietnam. As a province located in the Mekong Delta, An Giang develops mainly based on agriculture. Power supply services of An Giang Power Company still face with many problems such as: pressure of supply, power quality, continuity of power supply, attitudes of employees of service, delays in meet the requirements and feedback to customers. These problems stem from the following causes: the monopoly, lack of facilities and management skills, the inspection and supervision are not tight, power failures in an electricity network and so on. Thus, to improve business management, An Giang Power Company needs to understand clearly what factors that influence customer satisfaction.

Therefore, this study assesses the customer satisfaction with service quality in the case of An Giang Power Company to determine the customer service situation in order to improve service quality and ensure sustainable development for the An Giang Power Company.

## II. LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

### A. Customer satisfaction

There are many different perspectives assessment of the level of customer satisfaction. According to [17] suggested that customer satisfaction is an emotional reaction of customers responding to their experience with a product or service. [10] identified that the satisfaction as person's feeling of pleasure or disappointment which resulted from comparing a product's perceived performance or outcome against his/ her expectations. The satisfaction has three levels as follows: If customer perception is smaller than expected, the customer feels dissatisfied. If customers perceive by expectations, they feel satisfied. If the perception is greater than expected, the customer feels it is gratifying or enjoyable. However, in this study, power supply services relate to goods and products. In particular, electricity is one of the important things in our life. Therefore, study customer satisfaction for electric power supply service is an essential issue to help the managers of An Giang Power Company have an overview of the company's operation situation. Moreover, the article also provides useful information for managers to find appropriate improvement solutions to promote business activities. Thus, the quality of a service to be one of the most important factors that contribute to customer satisfaction.

### B. Service quality

Service quality is the most influential factor to customer satisfaction. The concept of service quality has been one of the most debated subjects in the services field. According to [23], quality of service reflected the interaction between customers and employees of service providers. [13] mentioned that service quality must be assessed on two aspects including service delivery process and results service. The author in [5] also suggested two areas of service quality that is technical quality and functional quality. The most common definition of service quality: Service quality is considered as the gap between service expectations and customer perceptions when using the service [20]. The service quality construct was measured by four indicators: reliability, responsiveness, assurance, and empathy.



### C. Relationship between service quality and customer satisfaction

The relationship between service quality and customer satisfaction is two different concepts but there is an intimate relationship with each other. The quality of services obtained is due to the comments and assessments of service users, while the satisfaction of customers using the service is brought from the quality of the service. Therefore, in order to improve customer satisfaction, service providers must improve service quality. In the study of the relationship between these two factors, [21] showed that service quality is the premise of customer satisfaction. [20] proposed a 5-component model of service quality (SERVQUAL), which is: reliability, service efficiency, facilities, service capacity and empathy. SERVQUAL is a reliable and widely used service quality measurement tool [7]. The initial scale [19] consists of ten components: reliability, responsiveness, competence, access, courtesy, communication, credibility, security, understanding/knowing the customer, and tangible.

[27] investigated service quality, customer value and satisfaction in the telecommunications industry in China. The authors point out that customer perceived service quality can influence a customer's behavioral intent directly by influencing customer value and customer satisfaction. The results showed that the four constituent service quality affects customer satisfaction include: tangible, reliability, guarantee, and quality network.

[17] investigated customer satisfaction in urban water supply in Nigeria. The study found that eight service quality components that affect customer satisfaction include: billing, reliability, pressure, helpfulness of staff, color of water, knowledge of staff, taste and courtesy of staff respectively. [25] studied service quality and its relationship with customer satisfaction by the SERVQUAL model in a gas company at Iran. Authors in [21] studied the quality of service in the power industry in India. The study concludes that seven components of service quality affect customer satisfaction including empathy, tangibility, reliability, assurance, responsiveness, security and stability.

Management Consulting Joint Stock Company (OCD) studied the satisfaction level of customers using electricity from 2013 to 2018 with customers of the Southern Power Corporation. The results show that quality service components that influence customer satisfaction including power supply, information to customers, electricity bill, customer service, business image; perception of electricity prices. [21] investigated the quality of service in the power industry in India. The results point out that the components of service quality affect customer satisfaction involving: trust, tangible, assurance, empathy, responsiveness, safety and stability. Therefore, this paper studies the customer satisfaction for the quality of power supply service in the case of An Giang Company. The components of this study is as follows: (1) Evaluation of power supply; (2) Information to customers; (3) Electricity bill; (4) Customer service; (5) Brand image.

Based on previous studies, the author proposes the following hypothesis (Fig.1):

H1. The evaluation of power supply service has effect on the customer's satisfaction.

H2. The information to customers has effect on the customer's satisfaction.

H3. The electricity bill has effect on the customer's satisfaction.

H4. The customer service has effect on the customer's satisfaction.

H5. The brand image has effect on the customer's satisfaction.

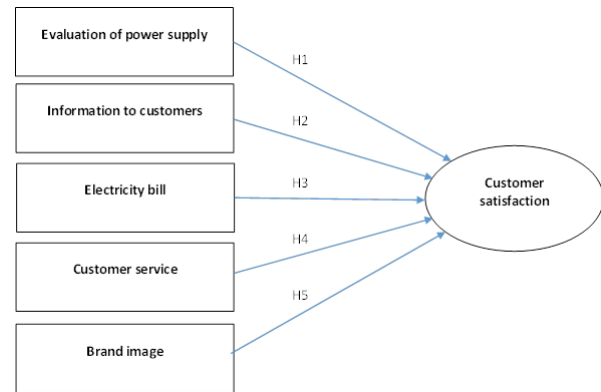


Fig.1. Research model

### III. METHODOLOGY

#### Sampling Method and Sample Size:

As the study is about measuring service quality of An Giang Power Company. In this study, 210 respondents of customers who using electricity in An Giang have been selected by using convenience sampling method. A survey was conducted in Long Xuyen city, Chau Doc city, Chau Phu district and Tan Chau town by using structured questionnaire. A convenience sampling process has been used to collect data for this study. After data collection, the correlation and multiple regressions analysis have been performed to test the strength of associations between variables by using SPSS software (25.0 versions). First, Cronbach's Alpha has been used to assess the reliability. In this research, the reliability coefficient Cronbach's alpha value for whole scale and each factor are over 0.6, which indicates that the measures used in this study are valid and highly reliable.

Then, exploratory factor analysis and multiple linear regression method were performed to test the research hypotheses.

TABLE I. SURVEY SAMPLE

Information		Number	Percentage (%)
1. Electricity users	Non-households	107	46,50
	Households	123	53,50
2. Electricity areas	Rural	143	62,20
	Urban	87	37,80

## IV. DATA ANALYSIS AND RESULTS

*Results of Cronbach's alpha*

The results of testing the Cronbach's Alpha reliability coefficient of the service quality components are greater than 0.6 (Table I). The lowest Cronbach's Alpha value is 0.651 (electricity bill) and the largest is 0.944 (customer service). The results show that these scales are reliable.

TABLE II. CRONBACH'S ALPHA RESULTS

Factors	Number of items	Cronbach's alpha
Evaluation of power supply (DG)	5	0,868
Information to customers (TT)	4	0,811
Electricity bill (HD)	4	0,651
Customer Service (DV)	8	0,944
Brand image (HA)	5	0,836

*Results of Exploratory Factor Analysis (EFA)*

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy varies between 0 and 1. In this study, the KMO value is 0,896 which point out that factor analysis is relevant and appropriate for this study (Table II). The Bartlett test ( $p=0.000<0.05$ ) also determined that the variables all display significant levels of correlation.

TABLE III: KMO AND BARTLETT'S TEST

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0,896
Bartlett's Test of Sphericity	Approx. Chi-Square	3561,492
	Df	210
	Sig.	0,000

TABLE IV. TOTAL VARIANCE EXPLAINED

Total Variance Explained							
Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings <sup>a</sup>
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	
1	9.175	43.689	43.689	8.892	42.342	42.342	7.695
2	2.127	10.130	53.819	1.853	8.823	51.165	5.651
3	1.681	8.003	61.822	1.124	5.353	56.518	5.786
4	1.183	5.631	67.453	.879	4.187	60.705	5.185
5	1.063	5.060	72.513	.652	3.106	63.811	3.678
6	.805	3.831	76.345				
7	.738	3.514	79.858				
8	.649	3.090	82.949				

Total Variance Explained							
Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings <sup>a</sup>
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	
9	.542	2.580	85.528				
10	.530	2.521	88.050				
11	.438	2.086	90.136				
12	.346	1.645	91.781				
13	.306	1.458	93.240				
14	.285	1.356	94.596				
15	.247	1.177	95.773				
16	.230	1.097	96.871				
17	.183	.870	97.740				
18	.165	.787	98.528				
19	.134	.639	99.167				
20	.106	.505	99.672				
21	.069	.328	100.000				

Extraction Method: Principal Axis Factoring.

TABLE V. COMPONENT MATRIX

Pattern Matrix <sup>a</sup>					
	Factor				
	1	2	3	4	5
DV6	.954				
DV1	.894				
DV4	.872				
DV7	.871				
DV5	.863				
DV3	.854				
DV8	.743				.224
TT2		.887			
TT3		.671			
TT1		.634			
TT4		.556			
HA3			.879		
HA5			.805		
HA1			.760		
DG1				.900	
DG2				.849	
DG5				.599	
HD2					.661
HD1					.539
HD4					.537
HD3					.508
<b>Eigenvalues</b>	<b>9,175</b>	<b>2,127</b>	<b>1,681</b>	<b>1,183</b>	<b>1,183</b>
<b>% Cumulative</b>	<b>42,342</b>	<b>51,165</b>	<b>56,518</b>	<b>60,705</b>	<b>60,705</b>

Extraction Method: Principal Axis Factoring.

Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Comparing the analytical results with the initially proposed theoretical model, the 5 components of service quality in power supply remain unchanged. The observed

variables still measure well for each component, there is no disturbance between observed variables of one component measured for another. However, the number of observed variables after analysis has decreased compared to the original scale. There are 5 observed variables which do not satisfy the factor load factor greater than 0.5, which has been excluded from the service quality scale, including: belonging to the Customer Service component (1 variable), Brand Image (2 variables) and Evaluation of Power supply (2 variables).

#### Results of Regression analysis

The results of the regression analysis show that the significance level of all independent variables is less than 0.05 (Table III). Therefore, the five components of service quality that affect customer satisfaction are statistically significant and the effect is positive (due to the positive regression coefficients). Moreover, the beta greater than 0 suggests that these factors have a strong impact on customer satisfaction. Brand image has the strongest impact on customer satisfaction (Beta = 0.294), followed evaluation of power supply (Beta = 0.237), electricity bill (Beta = 0.219), information to customers (Beta = 0.175) and customer service (Beta = 0.152).

TABLE VI. REGRESSION ANALYSIS RESULTS

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Variables	-0,516	0,237		-2,179	0,030		
	Evaluation of power supply	0,238	0,052	0,237	4,618	0,000	0,582	1,718
	Information to customers	0,215	0,065	0,175	3,295	0,001	0,542	1,843
	Electricity bill	0,219	0,044	0,219	4,978	0,000	0,794	1,260
	Customer Service	0,136	0,046	0,152	2,942	0,004	0,578	1,730
	Brand image	0,306	0,054	0,294	5,708	0,000	0,577	1,732

#### V. CONCLUSION

The research results show that 5 factors positively impact on customer satisfaction. This is reflected in the appropriateness of the hypotheses. The survey components include: (1) Evaluation of power supply; (2) Information to customers; (3) Electricity bill; (4) Customer service; (5) Brand image with 26 observed variables for 5 components. Most of the observed variables in the above components are rated by customers as being equal to 4 or more points for a 5-point scale (8 points for a 10-point scale). However, among the observed variables tested, 5 nonconforming variables were removed, the remaining is 21 observed variables. Based on the research results, An Giang Power Company has an overview to give specific solutions to improve service quality and improve business activities.

An Giang Power Company needs to focus on the group of brand image and evaluation of power supply. Because these factors have the strong impact on customer satisfaction. Researching and evaluating the satisfaction of customers using electricity of An Giang Power Company to determine

the current status of service provision is the basis for An Giang Power Company to have a correct view of the service. The company in specific is able to plan solutions to improve service quality, improve image and ensure the sustainable development of the company. Other power companies in general are also likely to learn from this case to enhance their operation as well as customer services in order to contribute to the implementation of local goals in improving the quality of life of rural and urban people, households and non-household customers.

This study only assessed customer satisfaction level of An Giang Power Company. Future studies can compare service quality between each department to help their know about strengths and weaknesses in order to improve service quality.

#### REFERENCES

- [1] Agyapong, G. K., "The effect of service quality on customer satisfaction in the utility industry-A case of Vodafone (Ghana)", *International Journal of Business and management*, 6(5), 2011, pp. 203-210.
- [2] Babakus, E., & Boller, G. W., "An empirical assessment of the SERVQUAL scale", *Journal of Business research*, 24(3), 1992, pp. 253-268.
- [3] Đặng Thanh, S & ctg., "Tax payers' satisfaction of about service quality of propaganda support service at Kien Giang Taxation Department", *Can Tho University Journal*, 2013, pp. 25.
- [4] Gerbing, D. W., & Anderson, J. C., "An updated paradigm for scale development incorporating unidimensionality and its assessment", *Journal of marketing research*, 25(2), 1988, pp.186-192.
- [5] Grönroos, C., "A service quality model and its marketing implications", *European Journal of marketing*, 18(4), 1984, pp.36-44.
- [6] Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C., *Multivariate Data Analysis* Prentice Hall. Upper Saddle River, NJ, 730, 1998.
- [7] Hemmasi, M., Strong, K. and Taylor, S., "Measuring service quality for planning and analysis in service firms", *Journal of Applied Business Research*, Vol. 10, No. 4, 1994, pp. 24-34.
- [8] Hoàng, T., & Chu, N. M. N., *Data analysis with SPSS*. Hồng Đức Publishing House, 2008.
- [9] Kiều Thị, H., "Research on customer satisfactory about service quality of Hai Au Hotel in Quy Nhon City – Binh Dinh Province", *Master Thesis, HCMC University of Economics*, 2011.
- [10] Kotler & Armstrong, *Marketing*, Grada Publishing, 2004.
- [11] Kotler, P., & Keller, K. L., *Marketing Management*, Pearson Prentice Hall, New Jersey, US, 2006.
- [12] Kotler, Philip, *Marketing Management*, Vietnam Education Publishing House Limited, 2003.
- [13] Lê Ngọc, S., "Research on customer satisfaction about service quality of public administration at Cu Chi People's Committee, HCMC", *Master Thesis, HCMC University of Economics*, 2011.
- [14] Lehtinen, U., & Lehtinen, J. R., "Service quality: a study of quality dimensions", *Service Management Institute*, 1982.
- [15] Nguyễn Đình, T., "Scientific research method in business, Phương pháp nghiên cứu khoa học trong kinh doanh", *Labour and Social Publisher Company Limited*, 2011.
- [16] Nguyễn Thị Mai, T., "Service quality, satisfaction, and loyalty of customer at supermarkets in HCMC", *Science and Technology Development Journal*, Volume 9, 10, 2006.
- [17] Ojo, V. O., "Customer satisfaction: A framework for assessing the service quality of urban water service providers in Abuja, Nigeria", (Doctoral dissertation, © Ojo, VO), 2011.
- [18] Oliver, R. L., *Satisfaction: A Behavioral Perspective on the Consumer*, New York: McGraw-Hill, 1997.

- [19] Parasuraman, A., Zeithaml, V. A., & Berry, L. L., "A conceptual model of service quality and its implications for future research", *Journal of marketing*, 49(4), 1985, pp. 41-50.
- [20] Parasuraman, A., Zeithaml, V. A., & Berry, L. L., "Servqual: A multiple-item scale for measuring consumer perc", *Journal of retailing*, 64(1), 1988, pp. 12.
- [21] Satapathy, S., Patel, S. K., Mahapatra, S. S., Beriha, G. S., & Biswas, A., "Service quality evaluation in electricity utility industry: an empirical study in India", *International Journal of Indian Culture and Business Management*, 5(1), 2011, pp. 59-75.
- [22] Spreng, R. A., & Mackoy, R. D., "An empirical examination of a model of perceived service quality and satisfaction", *Journal of retailing*, 72(2), 1996, pp. 201-214.
- [23] Survey data, "Customer satisfaction about service of Vietnam Electricity in cooperation between OCD Management Consulting Joint Stock Company during 2013 and 2018".
- [24] Svensson, G., "A triadic network approach to service quality", *Journal of Services Marketing*, 16(2), 2002, pp. 158-179.
- [25] Tabrizi, A.G., Ghayour, M. and Rajaei, Z., "Measurement of Service Quality and its Relationship with the Client's Satisfaction Through SERVQUAL Model in the Gas Company", *Middle-East Journal of Scientific Research*, 12 (8), 2012, pp. 1173-1181.
- [26] Trần Việt, H., "The affect of service quality in water supply to customer satisfaction: a case study in An Giang Power and Water Supply Joint Stock Company", Master thesis, HCMC University of Economics, 2013.
- [27] Wang, Y., Lo, H. P., & Yang, Y., "An integrated framework for service quality, customer value, satisfaction: Evidence from China's telecommunication industry", *Information systems frontiers*, 6(4), 2004, pp. 325-340.
- [28] Zeithaml VA & Bitner MJ, *Services marketing*. Boston: McGraw-Hill, 2000.

# Engineering Properties of Cement Mortar Produced with Mine Tailing as Fine Aggregate

Duy-Hai Vo

Department of Civil Engineering  
University of Technology and Education,  
The University of Danang  
Danang, Vietnam  
duyhai88@gmail.com

Khanh-Dung Tran Thi

Department of Civil and Construction  
Engineering  
National Taiwan University of Science and  
Technology  
Taipei, Taiwan  
khanhdungtran412@gmail.com

Mitiku Damtie Yehualaw

Faculty of Civil and Water Resource  
Engineering  
Bahir Dar Institute of Technology, Bahir  
Dar University  
Bahir Dar, Ethiopia  
mtkdmt2007@gmail.com

Chao-Lung Hwang

Department of Civil and Construction  
Engineering  
National Taiwan University of Science and  
Technology  
Taipei, Taiwan  
mikehwang@mail.ntust.edu.tw

Thi-My Ngo

Department of Civil Engineering  
University of Technology and Education,  
The University of Danang  
Danang, Vietnam  
myxdcn@gmail.com

Hoang-Anh Nguyen

Department of Civil Engineering, College  
of Engineering Technology, Can Tho  
University  
Can Tho, Viet Nam  
hoanganh@ctu.edu.vn

**Abstract—** This research investigated the feasibility of utilization of mine tailing, a disposal from mining exploitation process, to produce the cement mortar samples. Based on Densified Mixture Design Algorithm (DMDA) method, the mix-proportions in which fly ash (FA) accounted for 5% by total weight of FA and mine tailing were carried out on varied water-to-binder (w/b) ratios of 0.3, 0.4, and 0.5. Compressive strength, ultrasonic pulse velocity (UPV) and thermal conductivity tests were examined the engineering performance of mortar specimens. All mine tailing mortar specimens performed exceptional engineering properties, which were greatly affected by w/b ratio. The greatest compressive strength obtained in this study were 70.4 MPa in the mortar with w/b of 0.3. Increasing the w/b ratio had a negative impact on strength, UPV and thermal conductivity of mortar specimens.

**Keywords—** engineering performances, mine tailing, compressive strength, thermal conductivity.

## I. INTRODUCTION

The continuing development of China in infrastructure requires an increasing amount of construction materials. According to Cement, Concrete & Aggregate Markets in China, this demand is expected to grow by around 8% by 2027 [1]. That drives overexploitation of natural sand in some areas as well as a series of consequent environmental issues, for examples, increased river bed depth, water table lowering, intrusion of salinity and destruction of river embankment. Therefore, alternative materials should be well concerned to mitigate negative impacts from the problem. In addition, mining exploitation is harming the natural environment due to tailing disposal, which was commonly discharged into the adjacent rivers and seas [2]. Annually, 0.6 billion of iron ore tailings was disposed in China while the amount of the recycled tailing as resources was merely under 7% [3]. Also in this country, from 1949 to 2014, more than 30 billion tons of copper tailing was generated and most of them was dumped in ponds and probably caused serious environmental problems [4].

Utilizing mine waste tailing in construction could be an ideal solution for the two issues, which not only reuses the waste materials but also reduces natural aggregate consumption to reach the requirement for sustainable development. Numerous studies have investigated the performance of concrete and mortar prepared by abandoned tailing. Partial and complete replacement of natural aggregates with tailings exhibited outstanding mechanical strength and durability-related properties above the reference mixtures. Ismail and Al-Hashmi [5] found that partial replacement for sand by waste iron improved flexural strength and compressive strength of specimens containing waste-iron aggregate, significantly of the 20% waste-iron samples. Slump values, however, experienced an adverse trend with increasing content of waste-iron aggregates. The study of Ali Umara [6] found that the concrete with the substitution of ore tailing for sand of 25% achieved the highest compressive strength and splitting tensile strength throughout the curing period. This optimum number in an alike investigation of Zhong-xi Tian was 35%, from which the mechanical and durable performance of the concrete was closely comparable to that of natural-sand counterpart [7]. Thomas et al. [8], similarly, observed the higher compressive strength and flexural strength in the copper tailing concrete containing up to 60% substitution for natural sand. The incorporation of mine tailing, evidently, advanced the mechanical strength of hardened concrete and mortar. However, the mine tailing addition had a negative impact on flow-ability and shortened the setting time. Almost literature reviews mentioned mine tailing as sand partial replacement. Zhang et.al, on the other hand, investigated the optimal incorporation of copper tailing and manufactured sand in which copper sand was responsible for micronized sand in concrete [9]. The study indicated that the optimized gradation could be reached when 20% of manufactured sand was replaced by copper tailing, and the powder finer than 0.075 mm not only played a role as fillers in the void between the fine aggregate and cementitious materials in concrete but also served as micronized sand. This mixture performed the exceptional long-term compressive strength. In addition, the incorporation of fly ash as an additive



for tailing, likewise, has been studied to benefit both the cementation bonding and pore structure refinement of tailings, resulting the strength increase and permeability reduction [10].

This study presents the engineering performance of the mortars containing 100% mine tailing as fine aggregates produced with various w/b ratios. The mix-proportions were designed following to DMDA method, which was developed by Prof. Hwang in Taiwan. The mortar mixtures prepared with totally mine tailing aggregates and three w/b ratios: 0.3, 0.4, 0.5 were investigated. Compressive strength test, UPV test, and thermal conductivity test were conducted on the samples.

## II. MATERIALS AND EXPERIMENTAL METHODS

### A. Materials and mix design

To make the test mortar samples, mine tailings, sourced from China, were used as fine aggregates. Fig. 1 describes the particle size distribution of FA and tailing particles, which was conducted by using Mastersizer 2000 technique. As shown in Fig. 1, FA exhibited a fine material with a median diameter size D50 of 14.61  $\mu\text{m}$ , and specific surface area of 1.09  $\text{m}^2/\text{g}$ , while the mine tailing particles was slightly coarser with D50 of 22.359  $\mu\text{m}$  and specific surface area of 0.869  $\text{m}^2/\text{g}$ . Additionally, the water absorption ratio of tailing was approximately 2%.

TABLE I. CHEMICAL COMPOSITIONS OF RAW MATERIALS

Items	Tailing	FA	Cement
SiO <sub>2</sub>	71.76	63.9	22.01
Al <sub>2</sub> O <sub>3</sub>	9.02	20.2	5.57
Na <sub>2</sub> O	1.1	-	0.1
K <sub>2</sub> O	6.06	1.1	0.78
MgO	0.78	1.1	2.59
CaO	1.02	3.8	62.8
Fe <sub>2</sub> O <sub>3</sub>	3.33	6.5	3.44
SO <sub>3</sub>	2.91	1.4	2.08
Specific gravity (g/cm <sup>3</sup> )	2.73	2.2	3.15

The mixture proportions, designed by DMDA method developed by Prof. Hwang, are given in Table II. FA with the optimal 5% by total weight of tailing was used to fill into the voids of the tailing particles. The mixtures were prepared with various w/b ratios at 0.3, 0.4, and 0.5; superplasticizer (SP) was added to improve the workability of fresh mortar samples.

The chemical compositions of all materials are observed by using X-ray fluorescence analysis and are listed in Table I. Tailing and FA are mainly composed of SiO<sub>2</sub>, Al<sub>2</sub>O<sub>3</sub> and Fe<sub>2</sub>O<sub>3</sub>, while the OPC compositions mainly presented with SiO<sub>2</sub> and CaO. The binder used in the mortars were type I Ordinary Portland Cement (OPC) and class F fly ash, both are available in Taiwan. The mixing water used was local tap water. The superplasticizer was used to improve the workability of mortar samples. All the materials used conformed to the related ASTM standards.

TABLE II. MIX PROPORTIONS (KG/M<sup>3</sup>)

Mixtures	w/b	FA	Tailing	Cement	Water	SP
T30	0.3	67	1,274	680	194	30
T40	0.4	67	1,274	576	238	19
T50	0.5	67	1,274	497	276	6

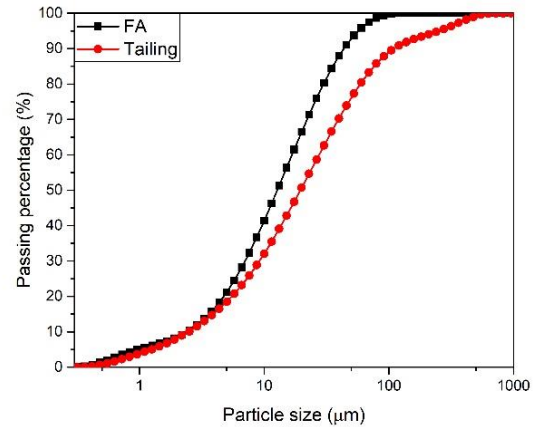


Fig.1. Particle size distribution of FA and Tailing

### B. Test programs

All the materials were prepared as Table II. Firstly, a part of water was poured into a mixer and then cement and FA were added and mixed in 3 mins to make a homogeneous paste. The tailing was added into mixer in 3 mins; then, the rest amount of water and SP were added to produce the fresh mortar. All the specimens were demoulded after 24 hours of casting and were cured in a chamber at  $25 \pm 2$  °C and humidity at  $95 \pm 5$  % until their testing age. For each mixture, three cubic samples were used to determine the compressive strength and ultrasonic pulse velocity test according to ASTM C109 and ASTM C597, respectively. Thermal conductivity was directly conducted by using the ISOMET 2104 device, which used a surface probe fitted with a temperature sensor to measure the thermal conductivity of mortar specimens within 10 mins. All the tests were conducted at ages 7 and 28 days.

## III. RESULTS AND DISCUSSIONS

### A. Compressive strength

The compressive strength data, collected on the 7<sup>th</sup> day and the 28<sup>th</sup> day of curing, improved with increasing curing time due to the compacted microstructure of the mortars [11]. Obviously, the strength of all of the mortars considerably developed to achieve from more than 40 MPa in T50 to over 70 MPa in T30 on the 28<sup>th</sup> day, making an increase up to 75%. The significant development of compressive strength in tailing-mixtures can be attributed to the fineness of tailings. According to Fall et.al [12], at a specific w/b ratio, in coarse and medium size tailing (15-35 wt% of 20  $\mu\text{m}$  for coarse tailings and >60 wt% of 20  $\mu\text{m}$  for medium tailing), the volume of void spaces between the tailing particles was sufficient to be filled by the hydration products. In addition, FA, moreover, as a filler in tailing, likewise, contributed to densify the structure of the mixture. Therefore, both filler effects fostered the compressive strength of mortar samples. The increasing w/b ratio resulted in the reduction of compressive strength. As shown in Fig. 2, T30 led among the

mixtures in the compressive strength while T50 performed the lowest values, both in initial stage and in later stage. The increasing the w/b ratio led to decrease the cement content and higher the water content, which significantly affected to the solid phase of hydration products and compressive strength of the mortar samples. Additionally, the compressive strength collected in T30 on the 28<sup>th</sup> day made a sharp increase by 75% compared to that on the 7<sup>th</sup> day, and was more exceptional than those of T40 and T50. These results agreed with the study implemented by Blesse et.al [8] and Sujing Zhao [13], in which increasing w/b was associated the strength loss. However, the bond between cement and aggregate interface was likely to enhance thanks to the rough and angular texture of tailing [6], and in consequence, benefited the strength development. Therefore, this could mitigate the strength loss in some extent.

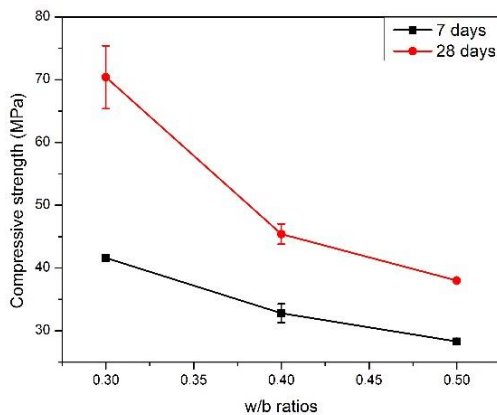


Fig. 2. Compressive strength of mortar samples

### B. Ultrasonic pulse velocity

The UPV values, recorded on the 7<sup>th</sup> day and the 28<sup>th</sup> day of curing, exhibited the gradual increase with curing time as shown in Fig. 3. In early ages, the UPV data ranged from approximately 3800 m/s to around 4100 m/s, being classified as Good condition, according to the suggested pulse velocity rating by Malhotra [14]. In later ages, these figures improved significantly to around 4200–4900 m/s in all mixtures, almost reaching the Excellent condition. This result is contributed by the hydration process of the raw materials along curing time, leading to the denser structure and higher UPV results of mortar specimens.

The result also revealed that the lower the w/b ratios, the higher the UPV values. As can be seen from the figure below, T30 achieved the highest UPV data while T50 performed the lowest values, both on day 7 and day 28 of curing. This record correlated to the corresponding compressive strength results. As mentioned above, the higher cement content and lower water content contributed to complete compaction microstructure of mortar samples, which increased the UPV results. On the other hand, the trend of UPV values can be associated with the fineness effect of tailing particles and the adequate w/b, which were advantageous to refine the pore matrix in the mortar and the sufficient hydration products to fill the voids between tailing particles [8, 15].

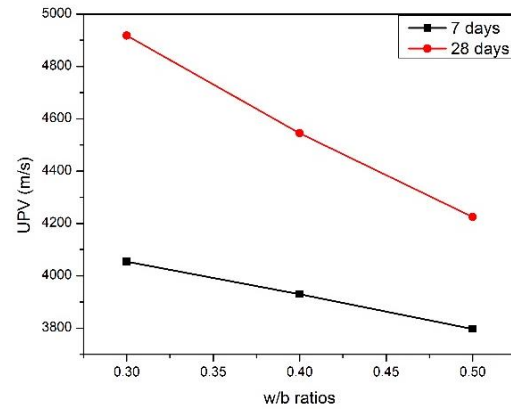


Fig.3. Ultrasonic pulse velocity mortar samples

### C. Thermal conductivity

Thermal conductivity reflects the ability to conduct heat in a material. The figure below describes thermal conductivity values of the three mortar samples during the testing period for 28 days. Owing to hydration processes, the hydration products produced along the curing time increased the mortar density, leading to the rise of thermal conductivity [16, 17]. On the other hand, increasing w/b ratio lowered the thermal conductivity in the mortar specimens. As shown in Fig. 5, T30 presented the highest thermal conductivity level while T50 conducted least heat among the three mortars. At 28-day of curing age, the thermal conductivity results were 1.01, 0.98 and 0.95 with the w/b ratios at 0.3, 0.4 and 0.5, respectively. This tendency could be attributed to the density of microstructure of the samples. The higher w/b ratio, the more water existed in fresh mortars and this water would evaporate and be replaced by air in hardened mortars. In addition, the lower cement proportion also affected to the hydration process of mortar specimens. Consequently, the porosity of mortar increased, causing the decrease in the ability of conducting heat.

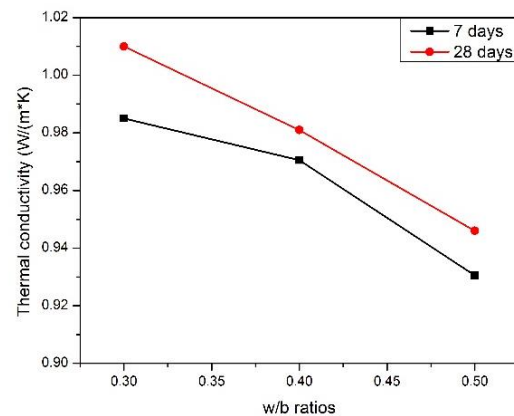


Fig. 4. Thermal conductivity mortar samples

## IV. CONCLUSIONS

This study exposed the possibility to utilize mine tailings as fine aggregate in the production of the mortar samples. It was pronounced that all the mine tailing mixture exhibited the good strength and engineering performances along curing time. Besides, the w/b ratio remarkably affected the engineering properties of the mortars. Increasing the w/b ratio apparently reduced the strength, UPV, and thermal conductivity. After 28 days of curing, mine tailing mortar

performed the high compressive strength range and good UPV result, which provided an attractive approach to utilize mine tailing as fine aggregates for building materials.

#### ACKNOWLEDGMENT

The authors gratefully acknowledge the Hwang's research group at the National Taiwan University of Science and Technology (NTUST) for assistance in conducting experimental works.

#### REFERENCES

- [1] Cement, Concrete & Aggregate Markets in China, AMID Co.
- [2] O. Onuaguluchi, Ö. Eren, Recycling of copper tailings as an additive in cement mortars, *Construction and Building Materials* 37 (2012) 723-727.
- [3] S. Zhao, J. Fan, W. Sun, Utilization of iron ore tailings as fine aggregate in ultra-high performance concrete, *Construction and Building Materials* 50 (2014) 540-548.
- [4] A. Khamseh, F. Shahbazi, S. Oustan, N. Najafi, N. Davatgar, Impact of tailings dam failure on spatial features of copper contamination (Mazraeh mine area, Iran), *Arabian Journal of Geosciences* 10(11) (2017) 244.
- [5] Z.Z. Ismail, E.A. Al-Hashmi, Reuse of waste iron as a partial replacement of sand in concrete, *Waste Management* 28(11) (2008) 2048-2053.
- [6] A.U. Shettima, M.W. Hussin, Y. Ahmad, J. Mirza, Evaluation of iron ore tailings as replacement for fine aggregate in concrete, *Construction and Building Materials* 120 (2016) 72-79.
- [7] Z.-x. Tian, Z. Zhao, C.-q. Dai, S.-j. Liu, Experimental Study on the Properties of Concrete Mixed with Iron Ore Tailings, 2016.
- [8] B.S. Thomas, A. Damare, R.C. Gupta, Strength and durability characteristics of copper tailing concrete, *Construction and Building Materials* 48 (2013) 894-900.
- [9] Y. Zhang, W. Shen, M. Wu, B. Shen, M. Li, G. Xu, B. Zhang, Q. Ding, X. Chen, Experimental study on the utilization of copper tailing as micronized sand to prepare high performance concrete, *Construction and Building Materials* 244 (2020) 118312.
- [10] J.K. Lee, J.Q. Shang, S. Jeong, Thermo-mechanical properties and microfabric of fly ash-stabilized gold tailings, *Journal of Hazardous Materials* 276 (2014) 323-331.
- [11] M. Fall, D. Adrien, J.C. Célestin, M. Pokharel, M. Touré, Saturated hydraulic conductivity of cemented paste backfill, *Minerals Engineering* 22(15) (2009) 1307-1317.
- [12] M. Fall, M. Benzaazoua, S. Ouellet, Experimental characterization of the influence of tailings fineness and density on the quality of cemented paste backfill, *Minerals Engineering* 18(1) (2005) 41-44.
- [13] S. Zhao, J. Fan, W. Sun, Utilization of iron ore tailings as fine aggregate in ultra-high performance concrete, *Construction and Building Materials* 50 (2014) 540-548.
- [14] V.M. Malhotra, Testing hardened concrete: nondestructive methods, Iowa State Press 1976.
- [15] X. Ke, X. Zhou, X. Wang, T. Wang, H. Hou, M. Zhou, Effect of tailings fineness on the pore structure development of cemented paste backfill, *Construction and Building Materials* 126 (2016) 345-350.
- [16] H. Uysal, R. Demirboğa, R. Şahin, R. Gül, The effects of different cement dosages, slumps, and pumice aggregate ratios on the thermal conductivity and density of concrete, *Cement and Concrete Research* 34(5) (2004) 845-848.
- [17] Q.L. Yu, P. Spiesz, H.J.H. Brouwers, Development of cement-based lightweight composites – Part 1: Mix design methodology and hardened properties, *Cement and Concrete Composites* 44 (2013) 17-29.

# Effect of Water-To-Solid Ratio on the Strength Development and Cracking Performance of Alkali-Activated Fine Slag under Water Curing Condition

Duy-Hai Vo

Department of Civil Engineering  
University of Technology and  
Education, The University of Danang  
Danang, Vietnam  
duyhai88@gmail.com

Khanh-Dung Tran Thi

Department of Civil and Construction  
Engineering  
National Taiwan University of Science  
and Technology  
Taipei, Taiwan  
khanhdungtran412@gmail.com

Mitiku Damtie Yehualaw

Faculty of Civil and Water Resource  
Engineering  
Bahir Dar Institute of Technology,  
Bahir Dar University  
Bahir Dar, Ethiopia  
mtkdm2007@gmail.com

Chao-Lung Hwang

Department of Civil and Construction  
Engineering  
National Taiwan University of Science  
and Technology  
Taipei, Taiwan  
mikehwang@mail.ntust.edu.tw

Hoang-Anh Nguyen

Department of Civil Engineering,  
College of Engineering Technology,  
Can Tho University  
Can Tho, Viet Nam  
hoanganh@ctu.edu.vn

Vu-An Tran

Department of Civil Engineering,  
College of Engineering Technology,  
Can Tho University  
Can Tho, Viet Nam  
tranvuan@ctu.edu.vn

**Abstract**—The present study aims to investigate the performance of alkali-activated fine slag under water curing condition. Ground granulated blast furnace slag (GGBFS) with a fineness of 6000 cm<sup>2</sup>/g was alkali-activated by the premixed alkaline solution from sodium hydroxide and sodium silicate liquid with the constant Na<sub>2</sub>O concentration at 5% and the modulus ratio of SiO<sub>2</sub>/Na<sub>2</sub>O at 0.5. Water-to-solid (w/s) ratio varied from 0.3, 0.4 and 0.5 to examine the effect of w/s ratio on the strength development and cracking performance of the AAS specimens. The compressive strength test, crack observation, and thermal conductivity test were conducted. The compressive strength results showed that this property was adversely affected by the w/s ratio, and the strength slowly increased after 7 curing days. Macro-cracks occurred in water-cured AAS cubes and they were more server in the specimens with lower w/s ratio. Besides, as the w/s ratio decreased, these AAS performed higher thermal conductivity.

**Keywords**—alkali-activated fine slag, cracking performance, thermal conductivity

## I. INTRODUCTION

Climate change and global warming because of CO<sub>2</sub> emission is currently one of the most important concerns worldwide [1]. Ordinary Portland cement (OPC) production is a crucial source of CO<sub>2</sub> emission [1, 2]. In 2018, the global OPC production reached 4.1 billion tons and this number is likely to continuously increase in the upcoming five-year period. This extremely large-scale production implies a proportional amount of CO<sub>2</sub> emitted into the atmosphere, which was reported to comprise of 5–7% of global total CO<sub>2</sub> emission [3], standing at approximately 0.73–0.85 tons of CO<sub>2</sub> released from the production of every ton of OPC [4]. Meanwhile, the main constituent of cement, clinker, consumes a large amount of natural raw materials (calcium carbonate (75–80 wt.%) and clays (20–25% wt.%) and a huge volume of natural fuels under very high temperature

process (≈1500°C) [5]. How to reduce the CO<sub>2</sub> emission and energy cost during cement production has become an urgent question for researchers to further explore.

Alkali-activated slag (AAS) is drawing growing attention as an alternative to OPC because of some advantages [6], including high strength at early ages, good resistance to chemical attack and chloride penetration, and low hydration heat [7], despite some remaining disadvantages known as fast setting time, large shrinkage, microcracks, possible occurrence of expansive reactions due to alkali-aggregate reactions, and higher formation of salt efflorescence [8]. Many factors regarding to properties of GGBFS or types of activators on the performance of AAS have been investigated [9, 10] while the concern about the effect of water-to-solid ratio (w/s) on this material is still limited. On the other hand, former research focused on AAS cured in air condition whereas the bath-cured companions have not yet received the adequate attention.

This study investigates the effect of w/s ratio on some engineering properties of AAS paste cured under water condition. Compressive strength, cracking performance, and thermal conductivity were examined in AAS paste with three w/s ratios: 0.3, 0.4, and 0.5.

## II. MATERIAL AND EXPERIMENTAL METHOD

### A. Materials and mix design

The raw material used in this study was fine GGBFS source with over 50% particles size smaller than 10μm. The X-ray fluorescence analysis, shown in Table I, was used to determine chemical compositions of the raw material. The main chemical compositions of GGBFS are SiO<sub>2</sub>, CaO, Al<sub>2</sub>O<sub>3</sub>, and MgO in total of 96% by weight, which mostly present in the glassy phase [11]. The Scanning electron microscopy (SEM) image of GGBFS showed that slag

particles was irregular shape as in Fig 1. Fig 2 illustrated the X-Ray Diffraction (XRD) pattern, showing GGBFS was nearly amorphous. The particle size distribution of the material is presented in Fig 3.

TABLE I. CHEMICAL COMPOSITIONS OF RAW MATERIALS

Items	GGBFS (6000) (%)
SiO <sub>2</sub>	36.7
Al <sub>2</sub> O <sub>3</sub>	14.67
Na <sub>2</sub> O	0.34
K <sub>2</sub> O	0.32
MgO	6.21
CaO	38.73
Fe <sub>2</sub> O <sub>3</sub>	0.32
SO <sub>3</sub>	1.67
Specific gravity (g/cm <sup>3</sup> )	2.92
Surface area (m <sup>2</sup> /g)	1.68
D50	8.83

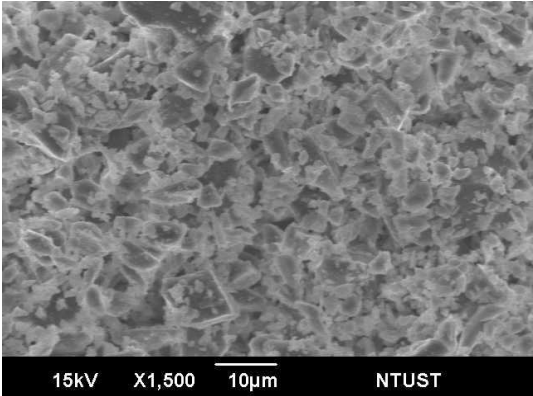


Fig. 1. SEM image of GGBFS

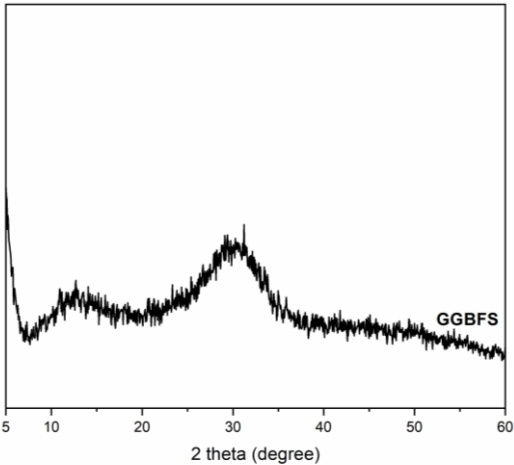


Fig. 2. XRD of GGBFS

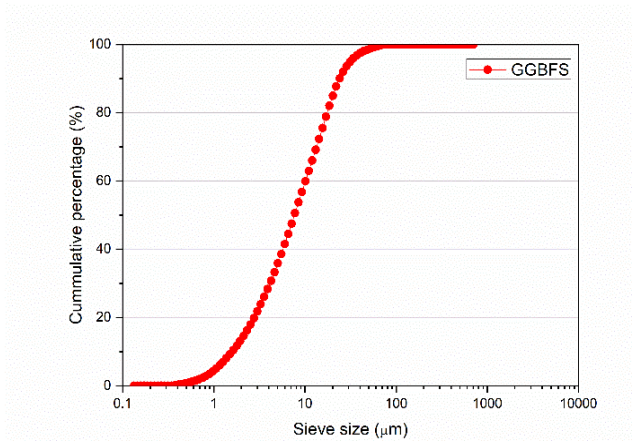


Fig. 3. Particle size distribution of GGBFS

The mixture proportions are given in Table II. The mixtures were prepared with various w/s ratios at 0.3, 0.4, and 0.5. The alkaline solution was a combination of sodium hydroxide (NaOH) with 98% purity and sodium silicate at Ms of 0.5 (Na<sub>2</sub>SiO<sub>3</sub>; with 25.7%-SiO<sub>2</sub>, 8.26%-Na<sub>2</sub>O, 66.04%-H<sub>2</sub>O, and SiO<sub>2</sub>/Na<sub>2</sub>O = 3.11). NaOH was supplied by Formosa Plastics Corporation in Taiwan, while the sodium silicate was imported from Pinnacle Industrial Co., Ltd. To avoid the exotherm from the dissolution of alkali, the dry NaOH pellets were first dissolved in water for 30 min and cooled down in room temperature in 24h before mixing with Na<sub>2</sub>SiO<sub>3</sub> solution for casting paste and mortar samples.

TABLE II. MIX PROPORTIONS (KG/M<sup>3</sup>)

Mixtures	w/s	NaOH (M)	GGBFS	Na <sub>2</sub> SiO <sub>3</sub>	Extra water
W-0.3	0.3	10	1618	157	228
W-0.4	0.4		1287	125	320
W-0.5	0.5		1069	104	381

Note: M- molarity

w- total water (water in NaOH and Na<sub>2</sub>SiO<sub>3</sub> solution and extra water); s- solid includes GGBFS, SiO<sub>2</sub> and Na<sub>2</sub>O in alkaline solution.

Firstly, a part of water was poured into the container of the mixer, then GGBFS 6000 was added. When the GGBFS was mixed homogeneously, the rest part of water was poured into the blended mixture. Subsequently, the prepared alkaline solution including NaOH solution and Na<sub>2</sub>SiO<sub>3</sub> solution was slowly added and mixed until the blend became homogeneous. These samples were cured in water curing condition to investigate the performance of AAS paste. The total time for mixing was around 6 mins.

B. Test programs

For each mixture, three cubic samples were used to determine the compressive strength at 1, 7 and 28 days of curing in according to ASTM C109. Thermal conductivity was directly measured by the ISOMET 2104 within 10 mins using a surface probe fitted with a temperature sensor. This test was conducted at day 7 and 28 of curing time. Cracking performance was observed until the 28<sup>th</sup> day. Time of the 1<sup>st</sup> crack occurred, crack length on average, the number of



cracks on average, the characteristic of cracks, and the number of cracked samples over the observed samples were monitored.

### III. RESULTS AND DISCUSSION

#### A. Compressive strength

Fig 4 describes the compressive strength development of mixtures cured water curing condition with 3 various w/s ratios: 0.3, 0.4, 0.5. As the w/s ratio increased, the compressive strength decreased correspondingly. It should be noticed that water acted as a carrier of alkalis and provided consistency to the fresh alkali-activated mixture [12]. However, high volume of water in the system caused more porosity, thus resulting in the strength decrease [13].

As can be seen from this figure, the compressive strength sharply increased in the early days and slowly rose in the later days of curing time. Particularly, the mixture of W-0.3 exhibited the 28<sup>th</sup>-day compressive strength lower than those on the 1<sup>st</sup> and the 7<sup>th</sup> day, which can be explained by significant cracks of W-0.3 mixture with curing time. In addition, mixtures of W-0.4 and W-0.5 exhibited the insignificant improvement in compressive strength result after 7-day of curing age. These both issues can be attributed to the cracks in paste matrix under water curing condition due to the excessive formation of hydration products [10]. The main hydration products of AAS paste are C-S-H gel and hydrotalcite-like phase (Ht). As mentioned in previous study [10], the high voluminous of Ht compared to C-S-H gel caused the expansion of AAS paste samples cured in water. The higher Ht formation led to unstable volume and more cracks performance of paste samples.

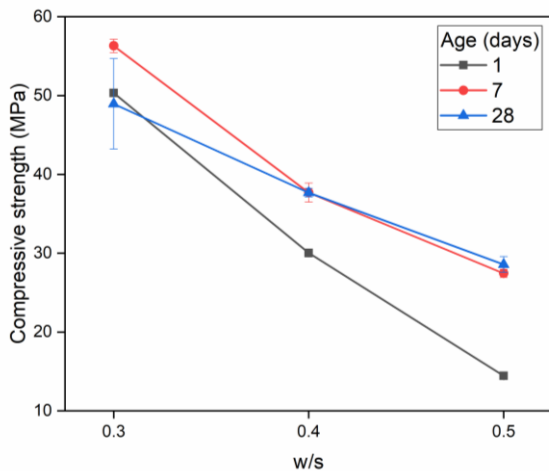





Fig. 4. Effect of w/s ratio on compressive strength

#### B. Cracking performance

Table III demonstrates cracking performance on the AAS specimens in bath curing condition. Crack phenomenon was investigated by observing the first day cracks occurred, the crack length average measured from whole 6 sides of all observed specimens, the average number of cracks, the characteristics of cracks, and the number of cracked samples over the total observed samples. It can be seen that cracking early occurred on the specimens and developed quite significantly during the curing time. Some of the cracks occurred on the corners of the samples, while the others occasionally arose in the center, which then could separate the samples into two pieces. This is likely because the finer

slag rapidly produced Ht, resulting in volume increase, more expansion, cracking, and eventually breaking [10]. The higher frequency of cracking occurrence was associated to the higher Ht generated in the paste.

TABLE III. CRACK OBSERVATION OF AAS PASTE SAMPLE UNDER WATER CURING CONDITION WITH DIFFERENT W/S

Mixtures	W-0.3	W-0.4	W-0.5
Sample's image			
Time the 1 <sup>st</sup> crack occurred	2 <sup>nd</sup> day	8 <sup>th</sup> day	No crack
Crack length average (cm)	57.00	29.33	0.00
No. of cracks on average	11.50	6.67	No crack
Characteristic	Cracks in group, continuing cracks on faces.	Continuing cracks on faces or curve cracks at the corners.	-
No. of cracked samples/observed samples	5/5	5/6	0/6

As shown in the table, the crack phenomenon increased with decreasing w/s ratio. The most serious cracks were observed in the samples with w/s = 0.3 when some paste cubes almost completely separated after 28 days. The time the 1<sup>st</sup> crack occurred in the mixture of W-0.3 was sooner than that in the mixture of W-0.4. Additionally, W-0.3 specimens also presented the highest crack length and number of cracks per sample. On the other hand, cracks did not appear in the samples with w/s = 0.5.

There was a hypothesis that when the hydration products were formed, they occupied the space of the water in paste matrix. In case of w/s = 0.3, the hydration process generated more products as well as the space dominated by water was less; thus, the hydration products not only replaced for water but also expanded the volume of the samples. Consequently, more cracks occurred.

#### C. Thermal conductivity

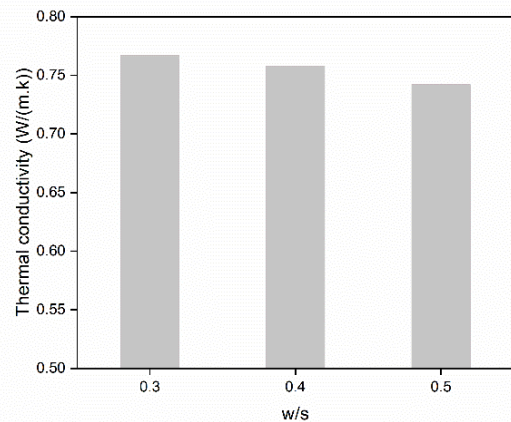


Fig. 5. Effect of w/s ratio on thermal conductivity

Thermal conductivity is also a nondestructive method to evaluate quantity of concrete, mortar, and paste. As the

matrix was denser, the thermal conductivity would be higher. Fig 5 illustrates thermal conductivity of the 3 AAS mixtures at the 28<sup>th</sup> day. As expected, the lower w/s ratio, the higher the thermal conductivity. The tendency of the thermal conductivity values corresponded to the aforementioned compressive strength. This result confirmed the study of Uysal et al. [14] which previously found that thermal conductivity of concrete was proportional to compressive strength.

#### IV. CONCLUSION

This study revealed the influence of w/s ratio on compressive strength, thermal conductivity, and cracking performance of water-cured AAS paste. In water curing condition, AAS paste exhibited a significant improvement in early strength, then slowed down or insignificantly developed on the later ages. Cracks occurred in the early age in bath-cured AAS pastes. Decreasing w/s ratio exacerbated cracks in AAS pastes cured in water condition.

#### ACKNOWLEDGMENT

The authors gratefully acknowledge the Hwang's research group at the National Taiwan University of Science and Technology (NTUST) for assistance in conducting experimental works

#### REFERENCES

- [1] M. Schneider, M. Romer, M. Tschudin, H. Bolio, Sustainable cement production—present and future, *Cement and Concrete Research* 41(7) (2011) 642-650.
- [2] E. Benhelal, G. Zahedi, E. Shamsaei, A. Bahadori, Global strategies and potentials to curb CO<sub>2</sub> emissions in cement industry, *Journal of Cleaner Production* 51 (2013) 142-161.
- [3] E. Gartner, Industrially interesting approaches to “low-CO<sub>2</sub>” cements, *Cement and Concrete Research* 34(9) (2004) 1489-1498.
- [4] A. Hasanbeigi, L. Price, E. Lin, Emerging energy-efficiency and CO<sub>2</sub> emission-reduction technologies for cement and concrete production: A technical review, *Renewable and Sustainable Energy Reviews* 16(8) (2012) 6220-6238.
- [5] R. Rehan, M. Nehdi, Carbon dioxide emissions and climate change: policy implications for the cement industry, *Environmental Science & Policy* 8(2) (2005) 105-114.
- [6] M. Torres-Carrasco, C. Rodríguez-Puertas, M.d.M. Alonso, F. Puertas, Alkali activated slag cements using waste glass as alternative activators. Rheological behaviour, *Boletín de la Sociedad Española de Cerámica y Vidrio* 54(2) (2015) 45-57.
- [7] S.A. Bernal, R. Mejía de Gutiérrez, A.L. Pedraza, J.L. Provis, E.D. Rodríguez, S. Delvasto, Effect of binder content on the performance of alkali-activated slag concretes, *Cement and Concrete Research* 41(1) (2011) 1-8.
- [8] A. Fernández-Jiménez, J.G. Palomo, F. Puertas, Alkali-activated slag mortars: Mechanical strength behaviour, *Cement and Concrete Research* 29(8) (1999) 1313-1321.
- [9] Z. Shi, C. Shi, S. Wan, Z. Ou, Effect of alkali dosage on alkali-silica reaction in sodium hydroxide activated slag mortars, *Construction and Building Materials* 143 (2017) 16-23.
- [10] C.-L. Hwang, D.-H. Vo, V.-A. Tran, M.D. Yehualaw, Effect of high MgO content on the performance of alkali-activated fine slag under water and air curing conditions, *Construction and Building Materials* 186 (2018) 503-513.
- [11] K. Gong, C.E. White, Impact of chemical variability of ground granulated blast-furnace slag on the phase formation in alkali-activated slag pastes, *Cement and Concrete Research* 89 (2016) 310-319.
- [12] J. Shekhovtsova, E.P. Kearsley, M. Kovtun, Effect of activator dosage, water-to-binder-solids ratio, temperature and duration of elevated temperature curing on the compressive strength of alkali-activated fly ash cement pastes %J *Journal of the South African Institution of Civil Engineering*, 56 (2014) 44-52.
- [13] S. Bernal, J. Provis, J. van Deventer, Impact of water content on the performance of alkali-activated slag concretes, *Durability of Concrete Structures*, Whittles Publishing, 2018, pp. 143-148.
- [14] M. Uysal, K. Yilmaz, Effect of mineral admixtures on properties of self-compacting concrete, *Cement and Concrete Composites* 33(7) (2011) 771-776.

# Determining Optimal Location and Sizing of STATCOM Based on PSO Algorithm and Designing Its Online ANFIS Controller for Power System Voltage Stability Enhancement

Huu Vinh Nguyen  
Ho Chi Minh City Power Corporation  
(EVNHCMC)  
Ho Chi Minh City, Vietnam  
nguyenhuvinhdlhcm@gmail.com

Hung Nguyen  
HUTECH Institute of Engineering  
Ho Chi Minh City University of  
Technology (HUTECH)  
Ho Chi Minh City, Viet Nam  
n.hung@hutech.edu.vn

Kim Hung Le  
The University of Danang, University of  
Science and Technology  
Da Nang Province, Viet Nam  
lekimhung@dut.udn.vn

Minh Tien Cao  
Faculty of Telecommunications  
Engineering  
Telecommunications University  
Nha Trang City, Viet Nam  
caominhtienkvtd@gmail.com

Tan Hung Nguyen  
Ho Chi Minh City Power Corporation  
(EVNHCMC)  
Ho Chi Minh City, Vietnam  
nguyentanhung1981@gmail.com

Tien Hoang Nguyen  
Ho Chi Minh City Power Corporation  
(EVNHCMC)  
Ho Chi Minh City, Vietnam  
nguyentienhoang21071995@gmail.com

Minh Vuong Le  
Ho Chi Minh City Power Corporation (EVNHCMC)  
Ho Chi Minh City, Vietnam  
vuongmle@gmail.com

**Abstract**—STATCOM is used to adjust bus voltage by injecting or absorbing the reactive power, therefore it is capable of improving power system voltage stability. However, it is important to find the optimal location and sizing of the device in a power system to support bus voltage in steady-state and design a controller to enhance transient voltage stability. In this paper, the optimal location and sizing of STATCOM are used based on the PSO method and the online ANFIS controller is designed for STATCOM to improve transient voltage stability under large disturbance. Online training of ANFIS is done using a neural identifier. And based on this identification, the weights and coefficients are adjusted timely. To demonstrate the performance of the proposed controllers, simulation results of the voltage response in time-domain are performed in MATLAB environment to evaluate the effectiveness of the designed controller for STATCOM.

**Keywords**—Static Synchronous Compensator (STATCOM), Particle Swarm Optimization (PSO), Artificial Neural Network (ANN), Adaptive Neuro-Fuzzy Inference System (ANFIS), ANFIS-Online, Voltage Stability.

## I. INTRODUCTION

Every Power Utilities has been struggling to enhance power quality and reliability for decades. In addition to the difficulties in the past, there are also a series of new difficulties such as the rapid development of renewable and distributed energy sources in which increase the share of non-linear loads. More and more known phenomenon could cause deterioration of supply voltage parameters, such as voltage fluctuations/flicker, voltage unbalance, higher harmonic content, voltage dips and swells, and interruptions in the power supply. These phenomena have a serious

negative impact on the units generating, transmitting, and distributing electricity, and on connected loads.

The modernization process has changed the electricity demand of consumers. Today, it is the users who have a higher demand for reliable and quality power supply. This higher demand is particularly serious for strategic customers such as financial institutions, hospitals, and military facilities, as well as industrial customers, where deterioration of supply voltage parameters can be a cause of the disruption to the fabrication process.

In order to improve power quality and reliability, mitigating power disturbances, voltage fluctuation, voltage drop, and harmonic distortion are interested in researching. One good solution to mitigating power quality disturbance is to install a Static Synchronous Compensator (STATCOM) [1]. For those Power Utilities (assume that they are solution investors), the optimal selection for the location and the capacity of the STATCOM device has always been a difficulty. This is not an easy task because the parameters of the power system heavily depend on the grid structure as well as the position of the load to be protected.

In this research, firstly the optimum STATCOM placement is done through the optimal power flow calculation. This is a type of grid network calculation that Power Utilities around the world are also very interested in and sought to solve for a long time. There have been many classic algorithms and modern techniques such as artificial intelligence could be applied to solve this problem such as Differential Evolution (DE) [2], Genetic Algorithm (GA) [3], Tabu Search (TS) [4] Ant Colony Optimization (ACO) [5-6] and so on.

Secondly, for the algorithms used to control the STATCOM, there have been many research papers giving methods to control STATCOM. Many of those research used a PID controller. However, in the real-life power network, it has many disturbance elements with complex configurations and its dynamic model is highly non-linear, the convention PID controller is concluded as not robust for STATCOM stability control. Recently, there are also several published papers that have supposed a new controller for STATCOM, such as [7], a fuzzy logic controller (FLC) has been used to enhance the power stability in the two-area four-generator interconnected power system. In another research, such as [8], a cooperating PI controller and FL controller to enhance the dynamic and steady-state performance of a speed controller based interior permanent magnet synchronous motor are presented. In the distributed network, a distributed STATCOM (DSTATCOM) is proposed and different fuzzy controllers with proportional plus integral control are implemented to control and maintain THD of the grid side current within the IEEE standards [9].

The PID controllers can be described by robust performances across a wide range of operating conditions and their functional simplicity. However, the high nonlinear of the power system means that a PID controller cannot perform well at all operating ranges, it can only be a robust performance at a particular operating range. The PID controllers can be basically divided into two categories. Firstly, the PID parameters are held throughout the control process; however, it is hard to meet good performances when the control system is nonlinear and heavily coupled. Secondly, in self-tuning PID, the parameters can be changed based on the estimation of the parameters [9-11]. In order to meet the optimal results, it is necessary to re-tune the PID controller when the operating range is changed, and other modern techniques from nonlinear control theory are required [11]. The neuro-adaptive learning techniques offer a procedure for the fuzzy modeling procedure to acquire information about a data set. The ANFIS control algorithm is also very well known for its robustness for nonlinear systems. Under uncertainty conditions, one good method to identify the model parameters of parallel manipulators is the ANFIS control algorithm. There are a lot of classic methods to train the model's fuzzy inference system in the ANFIS. However, in this case, in order to improve the training process of the ANFIS system, the online ANFIS controller is designed for STATCOM to improve transient voltage stability under large disturbance. For the purpose of showing this research paper results to enhance system stability, simulation results are presented to illustrate the effectiveness of the suggested controllers.

The contributions in this study are: (i) determine the position and capacity of STATCOM using the PSO algorithm with the components of the objective functions that are based on the total power loss in the power system, the penalties for the load buses whose voltage exceeds the range, and the penalties for branch power that exceed the allowable limit; (ii) improved STATCOM controller using ANFIS with online training technique and comparison with other controllers.

The paper is organized as below: Configuration of the studied system is introduced in the second section. The determining optimal location and sizing of STATCOM based on the PSO algorithm are presented in the third section. The

ANFIS online controller includes the online training method and anfis structure is introduced in the fourth section. Transient responses of the studied system between different controllers are presented in section five. Finally, specific important conclusions of this research are concluded in the sixth section.

## II. SYSTEM CONFIGURATION

### A. Studied System Configuration

The configuration of the studied system in this paper is the practical power network installed in Ho Chi Minh City, Vietnam that shown in Fig. 1. In the recent year, there are some disturbances that appear in ThuDuc power network, which is relevant to Intel bus, it made local Electricity of Viet Nam (EVN) had to report about these incidents to the power network, this is not appreciated for a power network because INTEL is a semiconductor producing company and it is very sensitive to the power disturbances or interruptions, the financial loss for INTEL, if interruptions happened at each time can be vast [12]. So, the requirements for power network stability are extremely strict in the current circumstance. For these reasons, the authors decided to research the issue of the ThuDuc power grid using Matlab software.

As shown in Fig. 1, The ThuDuc power transmission system has a power source 220 kV instead of the transmission line transferring the power from other 220 kV bus, five 250 MVA step-down transformers for reducing the voltage from 220 kV down to 110 kV, which is a loop circuit from ThuDuc bus to CatLai 220kV bus.

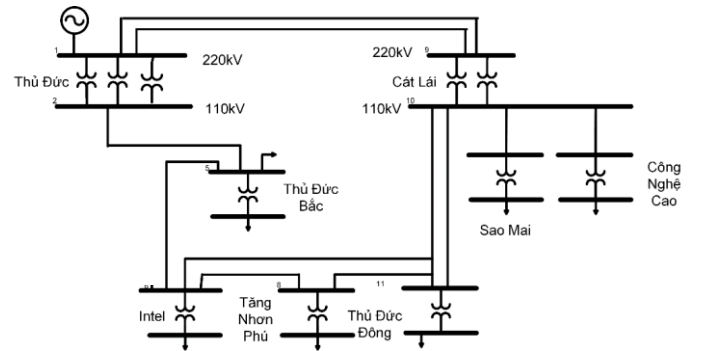


Fig. 1. Single line diagram of the ThuDuc power grid in Hochiminh city.

In the studied system, six load buses are called ThuDucBac, Intel, TangNhonPhu, ThuDucDong, SaoMai, CongNgheCao.

### B. STATCOM modeling

A STATCOM is designed for regulating the voltage at its terminals by compensating the amount of reactive power in or out from the power system. When the system voltage is low, the STATCOM injects the reactive power to the power system; when the voltage is high, it absorbs the reactive power. Besides, a STATCOM can be designed to act as an active filter to absorb system harmonics [13]. For analyzing the STATCOM, the mathematical model is used. In which, the output voltage is separated into two components represented in d and q axes as follow [14-15]:

$$v_{dsta} = V_{dsta} \cdot k_{msta} \cdot \sin(\theta_{bus} + \alpha_{sta}) \quad (1)$$

$$v_{qsta} = V_{dcsta} \cdot km_{sta} \cdot \cos(\theta_{bus} + \alpha_{sta}) \quad (2)$$

where:  $v_{dsta}$  and  $v_{qsta}$  are the voltages of d and q axes at the output terminals of the STATCOM, respectively;  $km_{sta}$  is the modulation index of the STATCOM;  $\alpha_{sta}$  is phase angle of the STATCOM;  $\theta_{bus}$  is the voltage phase angle of the common AC bus;  $V_{dcsta}$  is the DC voltage of the DC capacitor  $C_m$ . The relationship between DC voltage and current of the DC capacitor can be described as

$$(C_m)p(V_{dcsta}) = \omega_b[I_{dcsta} - (V_{dcsta}/R_m)] \quad (3)$$

In which:  $I_{dcsta}$  is the pu DC current flowing into the positive terminal of  $V_{dcsta}$ ;  $R_m$  is the pu equivalent resistance considering the equivalent electrical losses of the STATCOM;  $i_{qsta}$  and  $i_{dsta}$  are the currents in q and d axes flowing into the terminals of the STATCOM, respectively. The fundamental control block diagram of the STATCOM including the damping controller is shown in Figure 2. The DC voltage  $V_{dcsta}$  is controlled by the phase angle  $\alpha_{sta}$  while the voltage  $v_{sta}$  is varied by changing the modulation index  $km_{sta}$ .

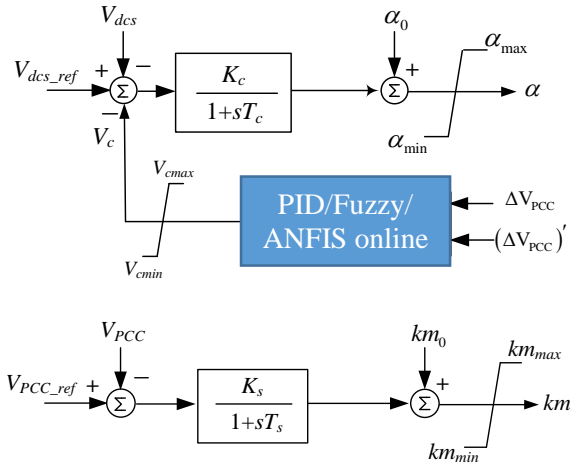


Fig. 2. The control scheme of STATCOM

### III. DETERMINING OPTIMAL LOCATION AND SIZING OF STATCOM BASED ON PSO ALGORITHM

#### A. Objective function

The purpose of determining optimal STATCOM placement is to improve the power quality of a system by minimizing the total cost function. Therefore, a multi-objective optimization problem is formed where its objective function includes sub-functions and operational constraints as follow:

$$Fit = \sum \Delta P_{line} + \alpha_v \sum_{j=1}^{N_d} (V_{lj} - V_{lj}^{lim})^2 + \alpha_s \sum_{j=1}^{N_{br}} (S_{lj} - S_{lj}^{lim})^2 \quad (4)$$

Where  $\alpha_v$ ,  $\alpha_s$  are the penalty multipliers for violated constraints of voltage and branch power.

In equation (4), the objective function depends on: (i) the total power loss in the power system under consideration, (ii) the sum of penalties for the load buses whose voltage exceeds the range, and (iii) the total of penalties for branch power that exceed the allowable limit.

#### B. Operation constrains:

For optimal placement and sizing of STATCOM, the following constraints are considered [16]:

##### 1) Power Flow Balance Equations

$$P_{gj} - P_{dj} = \sum_{k=1}^{N_{bus}} |V_j| |V_k| |Y_{j,k}| \cos(\theta_{j,k} - \delta_j + \delta_k) \quad (5)$$

$$Q_{gj} - Q_{dj} - Q_{sj} = \sum_{k=1}^{N_{bus}} |V_j| |V_k| |Y_{j,k}| \sin(\theta_{j,k} - \delta_j + \delta_k) \quad (6)$$

where  $P_{gj}$  and  $Q_{gj}$ , respectively, are the generated real and reactive powers, and  $P_{dj}$  and  $Q_{dj}$ , respectively, are the load real and reactive powers at bus  $j$ ;  $Y_{j,k}$  is the element of the bus admittance matrix.

##### 2) Size limit of STATCOM

The apparent power of a STATCOM as an active/reactive power compensator must not exceed its limiting value:

$$S_{stat} \leq S_{max} \quad (7)$$

##### 3) Security constrains

Bus voltages and branches power must be maintained within the limits of their nominal value given by

$$V_{lj,min} < V_{lj} < V_{lj,max}; j = 1 \dots N_d \quad (8)$$

$$S_l \leq S_{lmax}; l = 1 \dots N_l \quad (9)$$

$$S_l = \max\{S_{j,k}, S_{k,j}\}$$

#### C. Determining Optimal Location and Sizing of STATCOM Based on PSO Algorithm

##### 1) Particle swarm optimization (PSO) algorithm

PSO algorithm is a meta-heuristic one, which is inspired by the collective intelligence and behavior of birds and fish. This algorithm was first provided by Kennedy and Eberhart based on simple mathematical relations and considering the movement pattern of birds for the optimization of complex problems [17]. This algorithm starts to work by randomly creating an initial population (a group of particles). In fact, each particle shows a possible response. Each particle starts to move and search in the problem space in order to find the most appropriate point. In each step, this particle is fitted by its objective function and is placed toward the most appropriate direction to determine the most accurate and precise response. Each particle continues its movement each time using its experience and its neighbors in the problem



search space. Other particles move toward a particle with the best position and correct their directions. Therefore, the movement of particles in the problem search space depends on three factors including the present position of particle  $X_i^k$ , the best location that a particle has experienced (Pbest), and the best location that all of the particles have experienced (Gbest). In fact, in each cycle, the aim is to identify a particle that finds the best momentary position in the problem and enters the community with a new position, and the other particles move toward it considering the superiority of the most appropriate particle in terms of location. This cycle continues until all particles gather together at the best point [18]. These calculations are introduced based on Equations (10) and (11).

$$V_i^{(k+1)} = wV_i^{(k)} + c_1r_1(pb_{est_i} - X_i^{(k)}) + c_2r_2(gbest_i - X_i^{(k)}) \quad (10)$$

$$X_i^{(k+1)} = X_i^{(k)} + V_i^{(k)} \quad (11)$$

In (10),  $i = 1, \dots, N$ ,  $N$  is the population size (particle), and  $k = (1, 2, 3, \dots)$  is the iteration number in the algorithm process;  $V_i^{(k+1)}$  is the new velocity vector for the  $i^{\text{th}}$  particle;  $V_i^{(k)}$  indicates the existing velocity vector for the  $i^{\text{th}}$  particle;  $pb_{est_i}$  is the best position that the  $i^{\text{th}}$  particle has experienced;  $gbest$  is the best position that all particles have experienced. In (11),  $X_i^{(k)}$  is the present position of the  $i^{\text{th}}$  particle and the new position of the  $i^{\text{th}}$  particle;  $w$  is the weight inertia, which is used in the class of particles to ensure the convergence and is a suggestion in the range of 0.4-0.9;  $r_1$  and  $r_2$  are random numbers between 0 and 1;  $c_1$  and  $c_2$  are two fixed and positive values that are introduced as the personal learning factor and the global learning factor, respectively, and have a significant role in the algorithm's convergence controlling process. It is worth mentioning that the condition  $c_1 + c_2 \leq 4$  must always meet.

## 2) Determining the Optimal Location and Sizing of STATCOM

In this study, the optimal location and sizing of STATCOM are determined as followed (see Fig. 3):

- Step 1: Select parameters for PSO including the number of individuals  $N$ , number of iteration (bird\_step), personal learning factor and the global learning factor ( $c_1$ ,  $c_2$ ); the penalty multipliers for violated constraints ( $\alpha_v$ ,  $\alpha_s$ )
- Step 2: Calculation of power flow and fitness function according to equation (4)
- Step 3: Calculation and updating new velocity and position for each particle according to equations (10) and (11).
- Step 4: executed the loop from step 2 to step 3 until a certain termination condition is met.

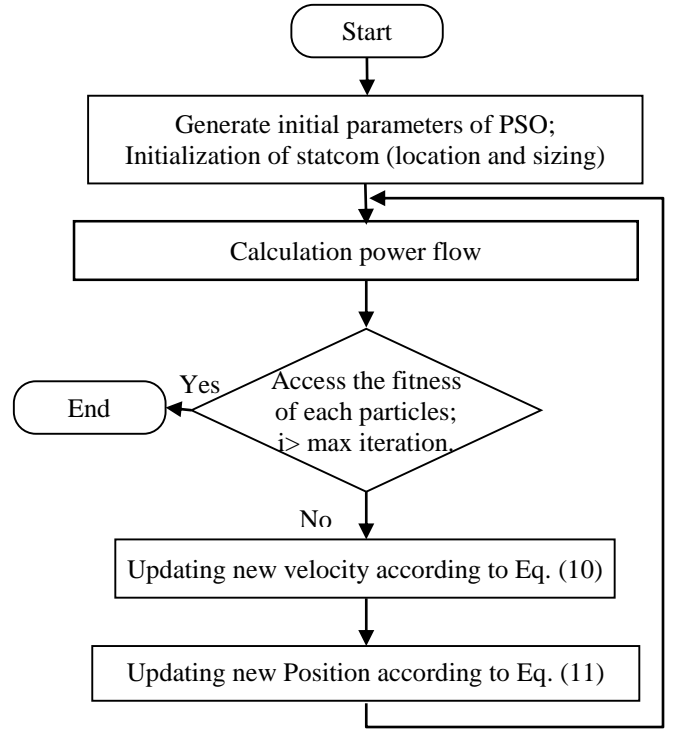


Fig. 3. Flowchart for determining location and sizing of STATCOM based on PSO

## IV. ANFIS-ONLINE CONTROLLER DESIGN FOR STATCOM

The ANFIS discriminates itself from normal fuzzy logic systems by the adaptive parameters, i.e., both the premise and consequent parameters are adjustable. The performance of the ANFIS system depends on their internal parameters, include of the membership function, the number of membership functions, their training data, and verify the number of data and training times that have to be carefully adjusted. The proposed intelligent ANFIS-online controller are presented in Fig 4. Five membership functions for both input error ( $e$ ) and its difference rate ( $de$ ) are applied.

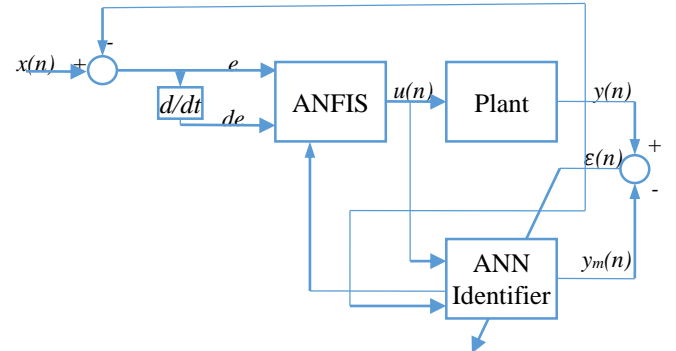


Fig. 4. Proposed ANFIS-Online controller

### A. ANFIS structure

A typical structure of used ANFIS is illustrated in Figure 4; in which a circle indicates a fixed node, whereas a square indicates an adaptive node. For simplicity, we consider two inputs  $x_1, x_2$ , and one output  $f$ . Among the fuzzy system models, the Sugeno fuzzy model is the most applicable cause of its high computational efficiency and interpretability, and integrated optimization, and adaptive techniques. In each model, the common rule set with two

fuzzy if-then rules can be explained as below [19]:

Rule  $i$ : if  $x$  is  $A_i$  and  $y$  is  $B_i$  then  $f_i = p_i x + q_i y + r_i$ .

where  $A_i$  and  $B_i$  are fuzzy sets in the antecedent and  $z=f(x, y)$  is a crisp function in the consequent;  $p_i, q_i, r_i$  are the updating parameters of rules.

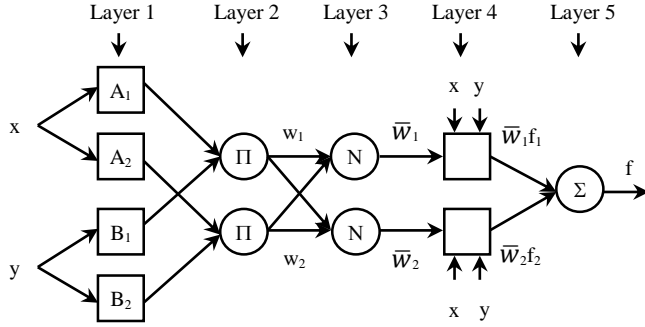


Fig. 5. Configuration of ANFIS

In this research, the ANFIS includes of five layers as follow [20]:

- **Layer 1:** In this layer input fuzzification takes place. Mathematically, this function can be explained as:

$$O_{ij}^{(1)} = \mu_j(I_{ij}^{(1)}) O_{ij}^{(1)} = \mu_j(I_{ij}^{(1)}) \quad (12)$$

where  $O_{ij}^{(1)}$  is the output of the Layer 1 node which corresponds to the  $j$ -th linguistic term of the  $i$ -th input variable  $I_{ij}^{(1)}$ .  $i$ -th is the quantities of input variable and  $j$  is the quantities of the linguistic term of each input. In this research,  $i = 2$  and  $j = 5$ .

$$\mu_j(I_{ij}^{(1)}) = e^{-\frac{1}{2} \left( \frac{x_i - c_{ij}}{\sigma_{ij}} \right)^2} \quad (13)$$

while the parameters  $\{\sigma_{ij}, c_{ij}\}$  are referred to as premise parameters or non-linear parameters and they adjust the shape and the location of the membership function. Those parameters are adjusted during the training mode of operation by the error back-propagation algorithm.

- **Layer 2:** In this layer, the total quantities of rules is 25. Each node output deputizes the activation level of a rule:

$$O_k^{(2)} = w_k \prod_{i=1}^q O_{ij}^{(1)} \quad (14)$$

$k$  is number rules.

- **Layer 3:** The  $k$ -th node's output is the firing strength of each rule divided by the total sum of the activation values of all the fuzzy rules. This leads to the normalization of the activation value for each fuzzy rule. This operation is simply expressed as:

$$O_k^{(3)} = \bar{w}_k = \frac{O_k^{(2)}}{\sum_{m=1}^y O_m^{(2)}} \quad (15)$$

- **Layer 4:** Each node  $k$  in this layer is accompanied by a set of adjustable parameters  $d_{1k}, d_{2k}, \dots, d_{N_{input}k}, d_{yk}, d_0$ , and implements the linear function:

$$O_k^{(4)} = \bar{w}_k f_k \quad (16)$$

$$O_k^{(4)} = (d_{1k} I_1^{(1)} + d_{2k} I_2^{(1)} + \dots + d_{N_{input}k} I_1^{(1)} + d_{0k})$$

- **Layer 5:** The single node in this layer calculates the overall output as the total of all incoming signals, which is written as:

$$O_k^{(5)} = \sum_{k=1}^y O_k^{(4)} = \sum_{k=1}^y \bar{w}_k f_k = \frac{\sum_{k=1}^y w_k f_k}{\sum_{k=1}^y w_k} \quad (17)$$

### B. Artificial neural network (ANN) Identifier

A Multilayer Perceptron (MLP) network is applied to represent the dynamics of the plant. The architecture of the MLP is illustrated in Figure 4. The proposed MLP network has 6 inputs, one hidden layer of 9 neurons with hyperbolic tangent functions, and an output layer with one neuron having linear node characteristics. The overall structure of the plant with the ANN identifier is shown in Figure. 3. The output of the ANN identifier is given by:

$$\Delta \hat{P}_s(n+1) = f(\Delta P_s(n), \Delta P_s(n-1), \Delta P_s(n-2), \dots, u(n), u(n-1), u(n-2)) \quad (18)$$

where  $\Delta P_s(n)$  is the power at the STATCOM at time step ( $n$ );  $u(k)$  is the control signal at time step ( $n$ );  $\Delta \hat{P}_s(n+1)$  is the output of the ANN Identifier, that is the forecasted at the time step ( $n+1$ ). The inputs of the ANN Identifier are standardized in the range of  $[-1, +1]$  before being employed in the neuron network. To rightly predict the output of the system at time step ( $n+1$ ), the identifier is first trained to carry out the estimated paradigm output,  $\Delta \hat{P}_s(n+1)$ , then it follows the actual plant output,  $\Delta P_s(n)$ , by minimizing the cost function is written as follow:

$$F(n) = \frac{1}{2} (\varepsilon(n))^2 = \frac{1}{2} (\Delta P_s(n) - \Delta \hat{P}_s(n))^2 \quad (19)$$

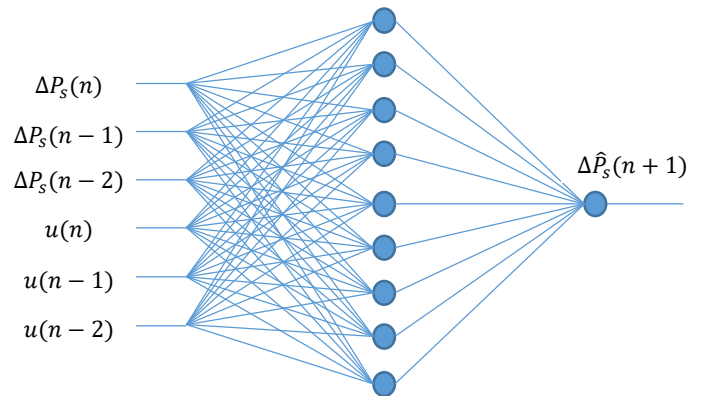


Fig. 6. ANN Identifier

The weights of the identifier are updated online using the gradient descent method as follows [21]:

$$W(n) = W(n-1) - \eta \nabla W J_i(n) \quad (20)$$

Where  $W(n)$  is the weights matrix at time  $n$ ,  $\eta$  is the network learning rate, and  $\nabla W J_i(n)$  is the gradient of  $J_i(n)$  with respect to the weight matrix  $W(n)$ . The gradient is calculated by:

$$\nabla W J_i(n) = -[\Delta P_s(n) - \Delta \hat{P}_s(n)] \frac{\Delta \hat{P}_s(n)}{\partial W(n)} \quad (21)$$

### C. Online training of ANFIS

The mission of the learning method in this architecture is to adjust all the adjustable weights such as Gaussian membership function variables and values of ANFIS rules which are called  $\{\sigma_{ij}, c_{ij}\}$ , and  $\{d_{2i}, d_{1i}, d_{0i}\}$ . The weight's modification is implemented to carry out the ANFIS output to match the training data. This research suggests a neuro-fuzzy control algorithm based on an artificial neural network identifier. Its topology is shown in Figure 5. The ANFIS weights are updated online using the output of the ANN identifier which be described above.

The error is defined as follows:

$$\varepsilon = x(n) - y_m(n) \quad (22)$$

The performance index for evaluation controller ability is defined as:

$$E(n) = \frac{1}{2}(\varepsilon(n))^2 = \frac{1}{2}(x(n) - y_m(n))^2 \quad (23)$$

To calculate  $\Delta\sigma = \sigma(n) - \sigma(n-1)$  and  $\Delta c = c(n) - c(n-1)$ , the authors use the below equations:

$$\sigma_{ij}(n+1) = \sigma_{ij}(n) + \left(-\frac{\partial E(n)}{\partial \sigma_{ij}}\right) \quad (24)$$

$$c_{ij}(n+1) = c_{ij}(n) + \left(-\frac{\partial E(n)}{\partial c_{ij}}\right) \quad (25)$$

To update the node of function parameters consist of  $\{d_{2i}, d_{1i}, d_{0i}\}$ , the authors use the below equations:

$$d_{2i}(n+1) = d_{2i}(n) + \left(-\frac{\partial E(n)}{\partial d_{2i}}\right) \quad (26)$$

$$d_{1i}(n+1) = d_{1i}(n) + \left(-\frac{\partial E(n)}{\partial d_{1i}}\right) \quad (27)$$

$$d_{0i}(n+1) = d_{0i}(n) + \left(-\frac{\partial E(n)}{\partial d_{0i}}\right) \quad (28)$$

In this research, the quantities of neurons, shown in Figure 4, are 5, 10, 20, 20, and 5 for layers 1, 2, 3, 4, and 5, respectively. This is explained that ten center weights in the membership functions (five for each input) and five consequent weights will be updated on-line. A system with these weights to update is considered relatively complex and time-consuming to calculate and train, especially when applied to real-time systems [22].

## V. SIMULATION RESULTS

### A. Determining Optimal Location and Sizing of STATCOM on ThuDuc power network

As the ThuDuc power network, that its single line diagram is drawn in Fig. 1, the 110kV Intel substation is supplied from three different power lines in from loops [12]. In detail, first power line is 110kV CatLai (171) – ThuDucDong - Intel (171), second power line is 110kV CatLai (172) – ThuDucDong -TangNhonPhu – Intel (172) and the third power line is 110kV ThuDuc (178) – ThuDucBac – Intel (176). The parameters of the PSO algorithm for determining optimal location and sizing of STATCOM on the ThuDuc power grid are chosen as in Table I. The comparison of all bus voltages in case of with and without STATCOM are shown in Table II. The optimal location and capacity of STATCOM are presented in Table III.

TABLE I. PSO PARAMETERS

PSO parameter	Value
Population Size	10
Maximum Number of Iterations	500
Inertia Weight (w)	1
Personal Learning Coefficient (c <sub>1</sub> )	1
Global Learning Coefficient (c <sub>2</sub> )	2
Inertia Weight Damping Ratio (w <sub>damp</sub> )	0.99

TABLE II. COMPARISON OF BUS VOLTAGE

Bus	Name of bus	V <sub>i</sub> (pu)	
		No STATCOM	STATCOM
1	220kV ThuDuc	1,0000	1,0000
2	220kV CatLai	0,9941	0,9982
3	110kV ThuDuc	0,9885	1,0002
4	110kV CatLai	0,9838	0,9974
5	ThuDucBac	0,9873	0,9998
<b>6</b>	<b>Intel</b>	<b>0,9804</b>	<b>1,0000</b>
7	TangNhonPhu	0,9792	0,9981
8	ThucDucDong	0,9799	0,9980
9	SaoMai	0,9802	0,9939
10	CongNgheCao	0,9948	0,9939

Calculation results show that the optimal location of STATCOM on the ThuDuc power network is bus 6, at Intel substation and its capacity is 0.15p.u. The total active power loss of the grid in case of without STATCOM is  $\sum \Delta P_{br(noSTAT)} = 0,266pu$ , this loss is reduced to  $\sum \Delta P_{br(STAT)} = 0,149pu$ , when installing STATCOM.

It is seem that, the lowest voltage in case of without STATCOM is at bus 7, TangNhonPhu substation,  $V_7 = 0.9792$  p.u. In case of install STATCOM at the optimal location, this voltage value is improved to  $V_{7(stat)} = 0.9981$  p.u, and all bus voltages are within the allowable range,  $V_i \geq 0,95pu$ .

TABLE III. LOCATION AND SIZE OF STATCOM

Optimal location	<b>Bus 6</b>
Total active power loss (pu)	0,149
Fit <sub>min</sub> (pu)	0,1491
Size of STATCOM(pu)	<b>0,15</b>
S <sub>cb</sub> (base power)	100MVA

### B. Simulation results of STATCOM controller based on ANFIS online

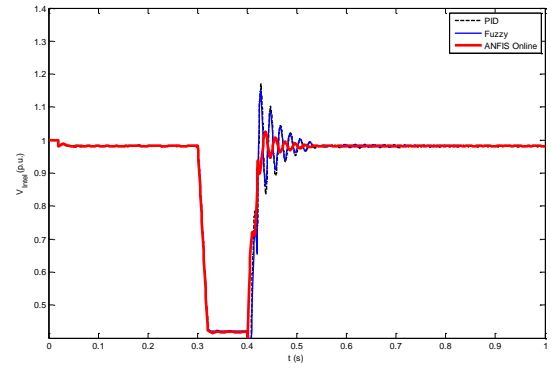
Based on the location of STATCOM, which is determined above, the authors apply the ANFIS-Online controller for voltage stability enhancement of the power system when a fault occurs.

For more clearly the effectiveness of the proposed controllers, Fig. 6 shows the simulation results of the studied system when a three-phase short circuit fault happened at the 110kV CongNgheCao substation. The voltage waves of buses that include Intel, ThuDucDong, CongNgheCao after a three-phase short circuit fault happened are presented in Figures 6 a) to d), respectively.

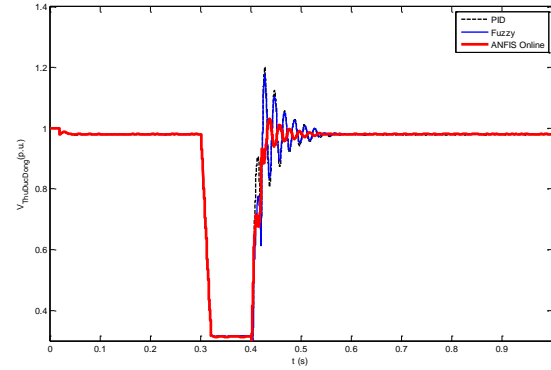
TABLE IV. COMPARISON OF CONTROLLERS

Items	Bus	PID	Fuzzy	ANFIS online
Settling time (s)	<b>Intel</b>	<b>0,487</b>	<b>0,486</b>	<b>0,457</b>
	ThuDucDong	0,507	0,489	0,457
	CongNgheCao	0,537	0,527	0,488
	TangNhonPhu	0,507	0,486	0,457
	ThuDucBac	0,447	0,447	0,437
	CatLai	0,527	0,527	0,477
Maximum Voltage (pu)	Intel	1,169	1,149	1,025
	ThuDucDong	1,201	1,176	1,031
	CongNgheCao	1,801	1,536	1,059
	TangNhonPhu	1,182	1,161	1,026
	ThuDucBac	1,065	1,059	1,007
	CatLai	1,662	1,401	1,056
Percent of Overshoot (POT) (%)	<b>Intel</b>	<b>19%</b>	<b>17%</b>	<b>4%</b>
	ThuDucDong	22%	19%	5%
	CongNgheCao	83%	56%	8%
	TangNhonPhu	20%	18%	4%
	ThuDucBac	8%	8%	2%
	CatLai	69%	42%	7%

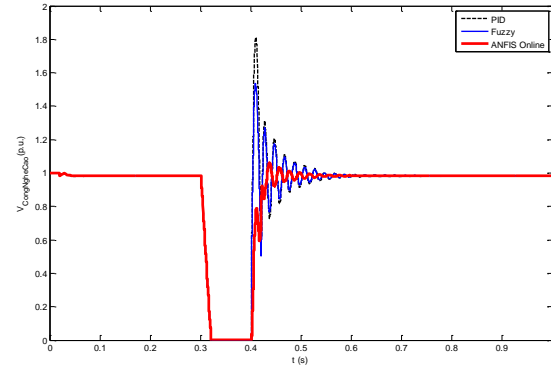
These figures describe the comparative transient responses of the studied system with the proposed STATCOM in cases of PID controller (black dotted lines), with Fuzzy controller (blue lines) and ANFIS-Online controller (red line) subject to a three-phase short-circuit fault at in 100ms (time for fault isolated) from 0.3s to 0.4s. With these figures, it is easily be seen that, by applying the Fuzzy and ANFIS online controllers for STATCOM device, the output value of voltage parameters are more stable and more effective, so that the voltage of each node is improved on overshoot and settling time after a three-phase short circuit fault occurred.



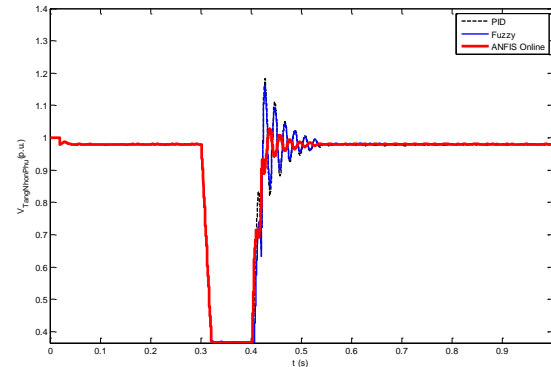
a) The voltage of Intel bus after a three-phase fault.



b) The voltage of ThuDucDong bus after a three-phase fault.



c) The voltage of CongNgheCao bus after a three-phase fault.



d) The voltage of TanNhonPhu bus after a three-phase fault.

Fig. 7. Voltage responses at each bus of the system when a three-phase short circuit fault happened in the CongNgheCao bus.

By observing the voltage response at the Intel bus that is shown in Fig. 6 c), it can be seen that the voltage of CongNgheCao bus drops down to zero during the fault

happened. The voltage response at ThuDucDong and Intel are also dropped to zero due to these buses are neighbors with CongNgheCao bus. However, with the response presented in Fig. 6 a), Fig. 6b), and Fig. 6b) the voltage magnitude of the Intel, ThuDucDong, TangNhonPhu bus are online drop to 0.3 p.u and 0.4 p.u, respectively. Based on the operation of the STATCOM and its designed controllers, a large amount of reactive power was supplied to improve the voltage of these buses.

In order to compare the efficiency between controllers, maximum voltage magnitude and percent of overshoot (POT) indexes and settling time are used. These indexes are shown in Table IV. In the case of applying the Fuzzy controller, the POT of voltages at Intel, ThuDucDong, CongNgheCao substations are 4%, 5%, and 8%, respectively. In the case of using the ANFIS-Online controller, the POT of voltages at Intel, ThuDucDong, CongNgheCao substations are 17%, 19%, and 56%, respectively. Meanwhile, the voltage overshoot in the case of the PID controller is 19%, 22%, and 83%. Comparison of the settling time of voltage after fault isolation, which is the time of voltage recovery within the permissible range of 5%, the ANFIS-Online controller gives the shortest time. With the ANFIS -Online controller, the settling time of voltage at Intel, ThuDucDong, CongNgheCao substations are 0.457s, 0.457s và 0.488s, respectively. Meanwhile, the voltage settling time in the case of the PID controller is 0.487s, 0.507s và 0.537s, respectively.

All the simulation results demonstrate that the ANFIS -Online is more effective than the Fuzzy controller and conventional PID controller.

## VI. CONCLUSIONS

This paper has presented the results of research on finding optimal location and sizing of STATCOM on ThuDuc power network using a particle swarm optimization algorithm. It can also be an application to another grid outside the scope of this research. The ANFIS-Online controller for STATCOM was designed and applied in the ThuDuc power network. STATCOM device can support fast response to the system to balance reactive power in the network help to improve dynamic voltage stability. The simulation results have shown that the ANFIS -Online controllers can be used to improve the system stability as well as the voltage quality more effective than conventional PID and Fuzzy controllers. This is due to the online training of ANFIS, that its the weights and coefficients of ANFIS are adjusted timely based on an artificial neural network identifier.

## REFERENCES

- [1] M. Farhoodnea, A. Mohamed, H. Shareef, and H. Zayandehroodi, "A Comprehensive Review of Optimization Techniques Applied for Placement and Sizing of Custom Power Devices in Distribution Networks," *PRZEGLĄD ELEKTROTECHNICZNY (Electrical Review)*, vol. 88, pp. 261-265, 2012.
- [2] Varadarajan M., "Optimal Power Flow Solution Using Differential Evolution," in Department of Electrical Engineering, Indian Institute of Technology, Madras, 2007.
- [3] Durairaj S., Kannan P.S., Devaraj D., "Application of Genetic Algorithm to Optimal Reactive Power Dispatch including Voltage Stability Constraint," *Journal of Energy & Environment*, vol. 4, pp. 63-73, 2005.
- [4] Abido M.A., "Optimal Power Flow Using Tabu Search Algorithm," *Electric Power Components and Systems*, vol. 30, pp. 469-483, 2002.
- [5] Boumediene Allaoua and Abdellah Laoufi, Collective, "Collective Intelligence for Optimal Power Flow Solution Using Ant Colony Optimization," *Leonardo Electronic Journal of Practices and Technologies*, vol. 12, pp. 88-105, 2008.
- [6] Aribi Fughar & Nwohu, M. N., "Optimal Location of STATCOM in Nigerian 330kv Network using Ant Colony Optimization Meta-Heuristic," *Global Journal of Researches in Engineering: F Electrical and Electronics Engineering*, vol. 14, no. 3, 2014.
- [7] A. Varshney and R. Garg, "Comparison of different topologies of the fuzzy logic controller to control D-STATCOM", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 2492-2497.
- [8] D. Shen, and P. W. Lehn, "Modeling, analysis and control of a current source inverter based STATCOM", *IEEE Trans. on Power Delivery*, Vol.17, No. 1, pp. 248-253, 2002.
- [9] A. Jain, K. Joshi, A. Behal, and N. Mohan, "Voltage regulation with STATCOMs: Modeling, control and results", *IEEE Trans. Power Delivery*, vol. 21, no. 2, pp. 726-735, 2006
- [10] A. Ganesh, R. Dahiya and G. K. Singh, "Development of simple technique for STATCOM for voltage regulation and power quality improvement," 2016 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Trivandrum, India, pp.1-6, 2016.
- [11] P. M. Anderson and A. A. Fouad, "Power System Control and Stability", IEEE Press, 2d ed., 2003.
- [12] EVNHCMC HVC, "Chương trình nâng cao độ tin cậy," 2018.
- [13] A. Ganesh, R. Dahiya, G. K. Singh, "Development of simple technique for STATCOM for voltage regulation and power quality improvement", 2016 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Trivandrum, India, December 14-17, 2016.
- [14] A. Varshney and R. Garg, "Comparison of different topologies of fuzzy logic controller to control D-STATCOM", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 2492-2497.
- [15] D. Shen, P. W. Lehn, "Modeling, analysis and control of a current source inverter based STATCOM", *IEEE Trans. on Power Delivery*, Vol. 17, No. 1, pp. 248-253, 2002.
- [16] Pisica I, Bulac C, Toma L, Eremia M (2009) Optimal SVC placement in electric power systems using a genetic algorithms based method. In: IEEE Bucharest power tech conference, pp 1-6
- [17] J. Kennedy, R. C. Eberhart, "Particle swarm optimization", *Proceedings of IEEE international conference on neural networks*, 1942-1948, New Jersey: IEEE Press, 1995.
- [18] P.S. You, "An efficient computational approach for railway booking problems", *European Journal of Operational Research*, Vol. 185, No. 2, pp. 811-824, 2008
- [19] H. V. Nguyen, M. T. Cao, H. Nguyen, and K. H. Le, "Performance Comparison between PSO and GA in Improving Dynamic Voltage Stability in ANFIS Controllers for STATCOM," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 4863-4869, 2019
- [20] M. Molinas, J.A. Suul, and T. Undeland, "Low voltage ride through of wind farms with cage generators: STATCOM versus SVC," *IEEE Trans. Power Electronics*, 2008, vol. 23, no. 3, pp. 1104-1117, 2008
- [21] P. Shamsollahi and O.P. Malik, "An Adaptive Power System Stabilizer Using OnLine Trained Neural Networks," *IEEE Transactions on Energy Conversion*, vol. 12, no. 4, pp. 382-387, 1997.
- [22] Miguel Ramirez-Gonzalez and O.P. Malik, "Simplified Fuzzy Logic Controller and its Application as a Power System Stabilizer," in *International Conference on Intelligent Applications on Power Systems*, 2009.



# An MCS–based Model to Qualify the Relationship between Worker’s Experience and Construction Productivity

Duy-Khanh Ha

Dept. of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
khanhhd@hcmute.edu.vn

Soo-Yong Kim

Dept. of Civil Engineering  
Pukyong National University  
Busan, South Korea  
kims@pknu.ac.kr

Van-Khoa Nguyen

Dept. of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
khoanv@hcmute.edu.vn

**Abstract**—Construction productivity is affected by many negative factors. Many previous studies have been conducted to investigate them and propose some solutions for productivity improvement. One of the main factors affecting construction productivity is a worker's skill and experience. Based on a survey with respondents in eight building projects, this study has found that formwork construction workers' average labor productivity is constant if their experience years are higher than 12 years. The actual productivity of formwork is higher than the baseline productivity. Eventually, a Monte Carlo simulation model has been developed by using a normal distribution and uniform distribution. The simulation results showed that the probability of obtaining the expected productivity is very high, with a standard error of only 2%.

**Keywords**—productivity, simulation, model, construction management, formwork

## I. INTRODUCTION

Productivity is a globally widespread problem (Naoum, 2016), even a buzz word (Jarkas and Bitar, 2012). As construction is a labor-intensive industry, productivity is typically related to laborers (Shehata and El-Gohary, 2011). The presence of labor with high productivity at each stage of a project's development plays a significant role in project success (Nasirzadeh and Nojedehe, 2013). In construction, labor productivity is critical to most construction projects (El-Gohary and Aziz, 2014). Even Jarkas and Bitar (2012) stated that labor cost comprises 30 to 50% of the overall project's cost and is regarded as a true reflection of the operation's economic success in most countries. In practice, many construction projects have a low labor productivity problem, which leads to project delay and cost. Therefore, over the last three decades, there have been many efforts to identify methods to enhance construction sites' current productivity (Naoum, 2016).

Working experience is one of the main characteristics of construction workers because it directly impacts the quality and quantity of outputs. It is the underlying reason why this study has been conducted. A case study related to the formwork task has been selected to demonstrate how the worker's experience affect construction productivity. The purposes of this study are: (1) to find out the average labor productivity of formwork construction for building projects, and (2) to develop a Monte Carlo-based model to qualify this productivity. The formwork activity that has been selected is wood formwork supported by steel bars.

## II. LITERATURE REVIEW

Over the years, many studies have been conducted to investigate critical factors affecting labor productivity in the construction industry. First, Hewage and Ruwanpura (2006) have indicated these factors for commercial construction projects include geographical dissimilarities, weather fluctuations, skill level variations in the workers, and job requests. Also, Mahamid (2013) proved that five of these factors for building projects involve rework, poor cooperation and message between project stakeholders, the economic status of the clients, inappropriate experience, and absence of essential materials. Besides, Hwang et al. (2017) also showed that five of these factors for green projects in the construction sector are workers' understanding, applied methods, design variations, labors' skill level, and arrangement, and progression of tasks. Most recently, Ahmed et al. (2018) found that skilled labor speeds up the construction schedule and productivity, thereby generating a very optimistic effect on the construction industry. Based on the above discussions, it could be concluded that workers have an essential role in productivity enhancement.

In general, construction labor productivity is affected by many factors. Modeling of construction labor productivity could be challenging when the effects of multiple factors are considered simultaneously (Sonmez and Rowings, 1998). One of the critical factors is related to the workforce because it is directly related to material handling missions (Sweis et al., 2008). Gatti et al. (2014) found a causal relationship between construction workforce physical strain and task level productivity. Besides, Loganathan and Kalidindi (2015) indicated that lack of training and improper workforce organizations, especially workers, has resulted in a sterile work environment. Also, regarding the labor factor, El-Gohary and Aziz (2014) revealed that two of the five significant factors affecting construction labor productivity in Egypt are labor experience and skills and competency of labor supervision. Furthermore, Pornthekasemsant and Charoenpornpattana (2015) proposed that five top factors influencing labor productivity in Thailand, ranked in descending order, are absenteeism of workers, low experience, financial shortage, inspection and instruction delay, and incomplete drawings. Furthermore, Choudhry and Zafar (2017) concluded that motivation, skills, and personal traits are three predictors for worker competency that affect construction productivity. This information demonstrates that

worker's characteristics play an essential role in creating high productivity.

Through a nationwide survey involving 1996 craft workers, Dai et al. (2009) found that workers do have a good understanding of the factors affecting their daily productivity, but the perceived productivity is relatively low. One of the keys to a successful construction productivity improvement program is a cost-effective method of obtaining accurate and consistent labor productivity (Noor, 1998). Productivity can be determined based on many different approaches. Therefore, it is calculated by dividing units of work placed or produced by person-hour (Shehata and El-Gohary, 2011) as below:

$$\text{Productivity} = \frac{\text{Units of work placed}}{\text{Man-hours used}} \quad (1)$$

As a result, labor productivity in construction is affected by both external factors and internal factors of a project. Most of them are deeply rooted in the poor management skills of engineers and managers. However, very few previous studies have been conducted to propose radical solutions to limit the probability of factors that cause low construction productivity. Besides, labor productivity is directly or indirectly related to other areas of construction management. Even it is difficult to measure construction productivity accurately because each project has unique characteristics.

### III. RESEARCH METHOD

Formwork is on-site construction activity. Therefore, the measurement of the productivity of this activity has been done through a work sampling sheet. The sheet consists of three sections: (1) brief information of project; (2) detailed information related to the construction performance of formwork activity; and (3) general information of in-charged engineers and observed workers. The personal information of engineers includes the number of experience years, academic degrees, and project stakeholders. The personal information of workers includes the number of experience years, level of training, and health status. The formwork activity productivity is calculated based on three following items: total area of formwork completed, time of start and finish, and numbers of workers involved. Thus, the unit of the actual general productivity (AGP) is defined as m<sup>2</sup>/hr, as in (1). Due to several specific difficulties, this study has used a convenient sampling method to collect data. To ensure random requirements when sampling, at least 10% of the daily formwork workers on the construction site in one working shift have been surveyed. The valid of the completed work sampling sheet received from the site have been checked based on two conditions. The first condition is that it must be filled fully by the in-charged engineer. The second condition is that the project manager must authenticate it. The collected data are then processed and analyzed by some appropriate statistical tools.

$$\text{AGP} = \frac{S}{W} \quad (2)$$

Where,

S = Total area of formwork completed (m<sup>2</sup>)

W = Number of working hours spent (hr)

According to the Vietnamese Quota No. 1776 (2007), the baseline productivity of the slab formwork ( $K_1$ ) is respectively 0.362 (m<sup>2</sup>/hr), 0.329 (m<sup>2</sup>/hr), and 0.301 (m<sup>2</sup>/hr) in accordance with at the height of less than or equal to 16m, less than or equal to 50m, and higher than 50m; meanwhile, that of the beam formwork ( $K_2$ ) is 0.333 (m<sup>2</sup>/hr), 0.301 (m<sup>2</sup>/hr), and 0.275 (m<sup>2</sup>/hr). The baseline general productivity (BGP) is then calculated based on the formula as below:

$$\text{BGP} = \frac{K_1 S_1 + K_2 S_2}{S} \quad (3)$$

Where,

$K_1$  = Baseline productivity of slab formwork (m<sup>2</sup>/hr)

$K_2$  = Baseline productivity of beam formwork (m<sup>2</sup>/hr)

$S_1$  = Area of slab formwork (m<sup>2</sup>)

$S_2$  = Area of beam formwork (m<sup>2</sup>)

For example,

If the height of a floor is less than 16m,  $S_1 = 21.600$  m<sup>2</sup>,  $S_2 = 1.290$  m<sup>2</sup>; thus  $S = 22.890$  m<sup>2</sup>, and  $\text{BGP} = (0.362 \times 21.600 + 0.333 \times 1.290) / 22.890 = 0.360$  m<sup>2</sup>/hr.

### IV. RESULTS

#### A. Properties of Workers

A total of 78 valid responses from 8 building projects have been collected after more than six months of the survey. The results of the descriptive statistical analysis for worker's properties are shown in Table I. The majority of them are people who have experienced between 5 and 10 years (41.0%), the next belongs to people who are between 10 and 15 years (32.1%), and the last belongs to people who are less than 5 years (16.6%), between 15 and 20 years, and higher than 20 years (1.3%). It would be better if the worker's percentage whose years of experience are between 15 and 20 years was increased.

TABLE I. PERSONAL PROPERTY OF FORMWORK WORKERS

Year of experience	Frequency	Percent
< 5 years	13	16.6
5 – 10 years	32	41.0
10 – 15 years	25	32.1
15 – 20 years	7	9.0
> 20 years	1	1.3
Total	78	100.0

#### B. The Mean of Productivity

The analysis results for the AGP and BGP of the formwork activity for slab and beam elements are shown in Table II. These results indicate that the average AGP and BGP are 0.388 m<sup>2</sup>/hr and 0.336 m<sup>2</sup>/hr. The deviation between AGP and BGP is 15.5%. It demonstrates that the actual performance of the formwork activity for slabs and beams is excellent. However, AGP's minimum value is less than that of BGP; whereas, the maximum value of AGP is higher than that of BGP. It means that the distribution of AGP is highly scattered. Based on Fig. 1, it is easily seen that most of the minimum values of AGP belong to workers who have years of experience of fewer than five years, and most of the maximum values of AGP belong to those who have years of experience

higher than nine years. Similarly, based on Fig. 1, the AGP is increased when the average number of experience years of workers is from 4 years to 12 years, and the AGP reach for a constant if the year of experience is almost higher than 12 years.

TABLE II. ACTUAL FORMWORK PRODUCTIVITY FOR ALL WORKERS

Productivity	Min	Max	Mean
AGP (m <sup>2</sup> /hr)	0.271	0.466	0.388
BGP (m <sup>2</sup> /hr)	0.302	0.381	0.336
Deviation (%)	15.5		

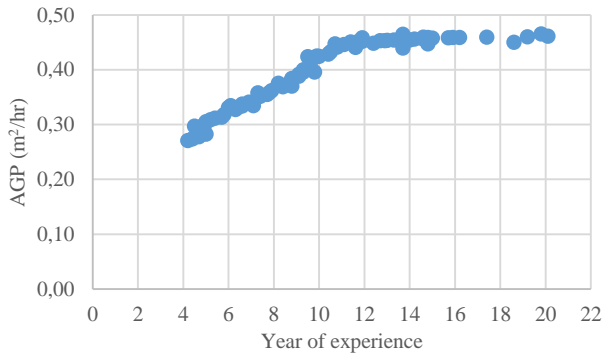


Fig. 1. Actual productivity of slab and beam formwork

### C. Relationship Model between Productivity and Worker's Experience

Productivity seems to be constant if the year of experience is higher than 12 years. Thus the relationship between AGP and worker experience (WE) is modeled for two cases as below:

*a) When a worker's experience is less than or equal to 12 years:* This study has also adopted the linear regression analysis to model the relationship between them. The method of variable selection for the analysis is an enter method. The analysis results show that the goodness of fit of the model is very high ( $R^2 = 98.1\%$ ). In addition, there is no statistical difference between regression and residual because the significance level is 0.000 less than 0.05. Moreover, the variance inflation factor (VIF) of 1.0 indicates that the collinearity does not happen (see Table III). Based on the results described in Table IV, the relationship between productivity and worker's experience can be expressed as (3). The standard errors are tiny. There is no difference between actual values and predicted values because their significance level is 0.000 (see Fig. 2).

$$AGP = 0.024 * WE + 0.177 \quad (3)$$

TABLE III. MODEL SUMMARY

Model	N	R Square	Std. Error of the Estimate	Sig.	VIF
1	56	0.981	0.00805	0.000	1.0

TABLE IV. COEFFICIENTS

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	0.177	0.004		48.181	0.000
	Worker's Experience	0.024	0.000	0.990	52.578	0.000

Dependent Variable: Productivity

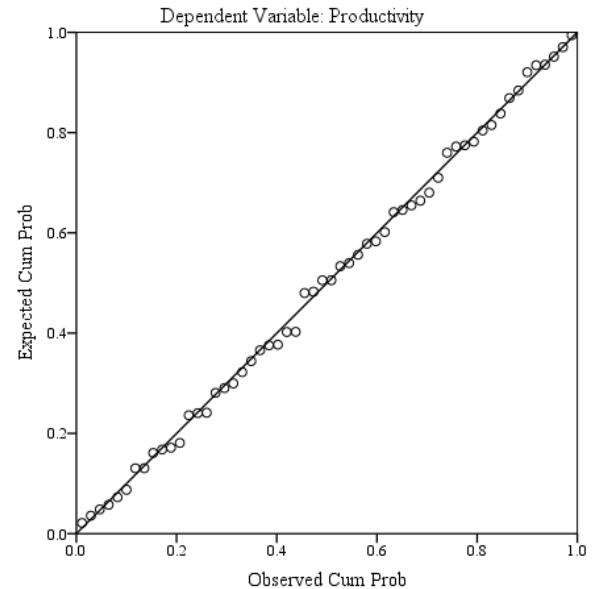


Fig. 2. Regression standardized residual

*b) When a worker's experience is higher than 12 years:* The descriptive analysis results show that the AGP is higher than the BGP with a deviation of 35.7% (see Table V). It indicates that the current practice of productivity performance is excellent.

TABLE V. ACTUAL FORMWORK PRODUCTIVITY FOR WORKERS WHO ARE 12 YEARS OF EXPERIENCE OR ABOVE

Productivity	Min	Max	Mean
AGP (m <sup>2</sup> /hr)	0.440	0.466	0.456
BGP (m <sup>2</sup> /hr)	0.302	0.381	0.336
Deviation (%)	35.7		

### D. Monte Carlo Simulation

The Monte Carlo Simulation (MCS) is a useful tool for qualifying the impact of a change in the independent variables on the dependent variables. In this study, these variables are the worker's experience (WE) and actual general productivity (AGP). One of the essential steps to perform the MCS is to define the probability distribution of the variables. Therefore, this study has adopted the Kolmogorov-Smirnov (K-S) test. In statistics, the K-S test is a nonparametric test of the equality of a continuous or discontinuous variable. If used to compare a sample with a reference probability distribution, it is called the one-sample K-S test. Previous studies stated that it is very suitable for small sets of data. In the test, three reference distributions have been selected to check the distribution of the collected data, including Normal, Uniform, and

Exponential. It is kindly noted that the Poisson distribution has not been considered because the data of this study are not integers. This test's statistical hypothesis can be stated as the following: There is no difference in distribution between the collected data and the reference data at the significance level of 0.05. The test results for the variables mentioned in (3) are shown in Table VI, Table VII, and Table VII. It proves that the normal distribution and uniform distribution are suitable to run the MCS because their asymptotic significance, respectively 0.579 and 0.316, is higher than 0.05. Whereas, the exponential distribution is rejected because its asymptotic significance is less than 0.05.

TABLE VI. TEST RESULTS FOR NORMAL DISTRIBUTION

Statistics		WE
Normal Parameters	Mean	7.714
	Std. Deviation	2.39
Most Extreme Differences	Absolute	0.104
	Positive	0.104
	Negative	-0.076
Kolmogorov-Smirnov Z		0.779
Asymp. Sig. (2-tailed)		0.579

TABLE VII. TEST RESULTS FOR UNIFORM DISTRIBUTION

Statistics		WE
Uniform Parameters	Minimum	4.20
	Maximum	11.90
Most Extreme Differences	Absolute	0.128
	Positive	0.128
	Negative	-0.036
Kolmogorov-Smirnov Z		0.960
Asymp. Sig. (2-tailed)		0.316

TABLE VIII. TEST RESULTS FOR EXPONENTIAL DISTRIBUTION

Statistics		WE
Exponential parameters	Mean	7.714
Most Extreme Differences	Absolute	.420
	Positive	.214
	Negative	-.420
Kolmogorov-Smirnov Z		3.142
Asymp. Sig. (2-tailed)		.000

This study has run the MCS with 1,000 trials based on the normal distribution and uniform distribution. The necessary values used in the trials have been selected randomly. The confidence level used for prediction control is 95%. The MSC results at the certainty level of 50% based on the normal distribution and uniform distribution are described in Fig. 3, and Fig. 4. For the normal distribution, the results show that the certainty level is 50% if the range of WE is from 0.322 m<sup>2</sup>/hr to 0.399 m<sup>2</sup>/hr. In addition, the entire range of WE is between 0.145 m<sup>2</sup>/hr and 0.585 m<sup>2</sup>/hr. After 1,000 trials, the standard error of the mean is only 0.2%. Similarly, for the uniform distribution, the results of the simulation can be summarized as the following: certainty level is 50%, certainty range is from 0.325 m<sup>2</sup>/hr to 0.417 m<sup>2</sup>/hr, the entire range is between 0.278 m<sup>2</sup>/hr and 0.462 m<sup>2</sup>/hr, and standard error of

the mean is also 0.2%. The trend of certainty in catching the AGP is also presented in Fig. 5 and Fig. 6 for normal distribution and uniform distribution, respectively.

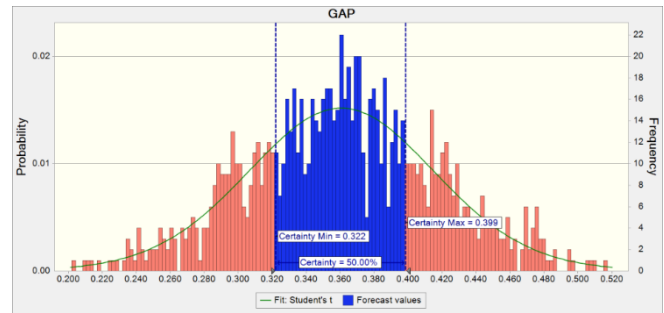


Fig. 3. Probability of AGP based on a normal distribution of WE

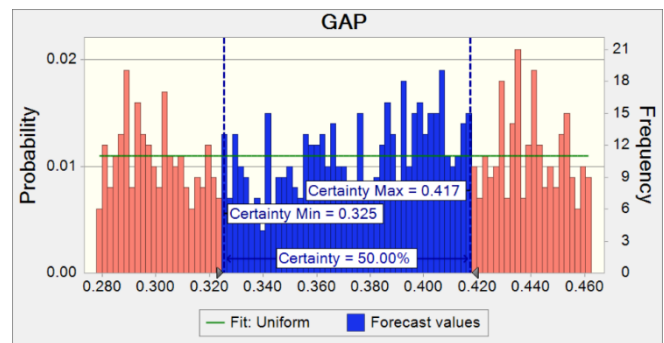


Fig. 4. Probability of AGP based on a uniform distribution of WE

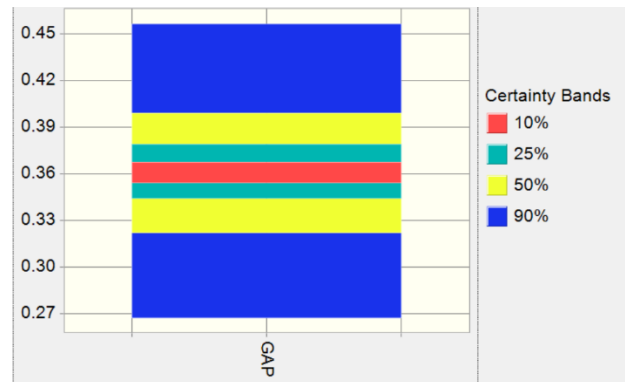


Fig. 5. The trend of certainty of AGP based on a normal distribution of WE

Although the accuracy of the simulation is good, the statistics of the forecast values need to be checked. The purpose is to validate the distribution used for the probability. The result of the check shows that these statistics are reasonable (see Table IX). For example, the Skewness statistic of 0.003 is very close to 0, and the Kurtosis statistic of 3.17 is very near 3. However, the mean value of predicted AGP based on the normal distribution is 0.361 m<sup>2</sup>/hr different from that of the uniform distribution of 0.373 m<sup>2</sup>/hr. As early mentioned, the MCS's main purpose is to define the probability of the value of AGP under the changes of WE. First, the results of the simulation (see Table X) prove that the probability of gaining each of the AGPs is different if it is based on the normal distribution and uniform distribution. Second, the range between the minimum value and the

maximum value is 0.439 for normal distribution; meanwhile, it is only 0.184 for uniform distribution. In general, the accuracy of these two simulations is very high for the prediction.

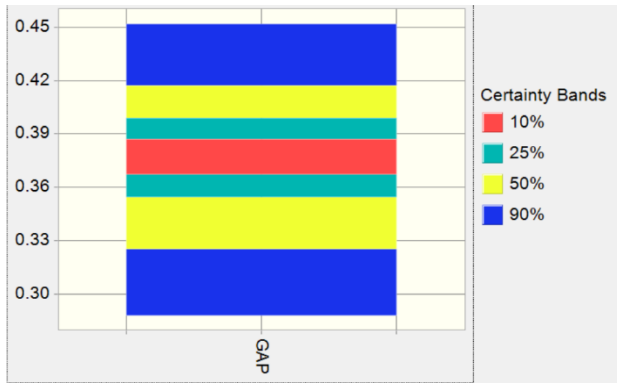


Fig. 6. The trend of certainty of AGP based on a uniform distribution of WE

TABLE IX. STATISTICS OF FORECAST VALUES

Statistics	Forecast Values	
	Normal	Uniform
Trials	1,000	1,000
Base Case	0.362	0.362
Mean	0.361	0.373
Median	0.361	0.378
Mode	---	---
Standard Deviation	0.057	0.053
Variance	0.003	0.003
Skewness	-0.0243	-0.1366
Kurtosis	3.17	1.79
Coeff. of Variation	0.1592	0.1429
Minimum	0.145	0.278
Maximum	0.585	0.462
Range Width	0.439	0.184
Mean Std. Error	0.002	0.002

## V. CONCLUSIONS

In construction, labor productivity is affected by the worker's experience. The higher the worker's experience, the higher the productivity. This study has identified the average labor productivity through a survey with 78 construction workers with different working experience in formwork activity. In detail, when a worker's experience years are less than or equal to 12 years, there is a linear relationship between it and productivity. This relationship is fitted with an accuracy of 98.1%. In addition, when a worker's experience years are higher than 12 years, the average labor productivity is almost constant with 0.456 m<sup>2</sup>/hr. In general, the actual labor productivity is higher than the baseline labor productivity as the worker's experience years are higher than seven years. Based on MCS simulation, this study has qualified the average labor productivity by using two data distributions, i.e., normal distribution and uniform distribution. The simulation's main

results show that the general average productivity (AGP) can be gained according to its probability.

TABLE X. PROBABILITY OF PREDICTION FOR AGP

Percentiles	Forecast Values	
	Normal	Uniform
0%	0.145	0.278
10%	0.289	0.296
20%	0.312	0.315
30%	0.331	0.336
40%	0.348	0.359
50%	0.361	0.378
60%	0.373	0.395
70%	0.391	0.409
80%	0.410	0.427
90%	0.433	0.442
100%	0.585	0.462

## ACKNOWLEDGMENT

We would like to thank all the respondents in the survey of this study for their help. We would also like to express our gratitude to grant support from the Ho Chi Minh City University of Technology and Education.

## REFERENCES

- [1] M. E. Shehata and K.M. El-Gohary, "Towards improving construction labor productivity and projects' performance," *Alexandria Engineering Journal*, vol. 50, no. 4, pp. 321–330, 2011.
- [2] S. G. Naoum, "Factors influencing labor productivity on construction sites: A state-of-the-art literature review and a survey," *International Journal of Productivity and Performance Management*, vol. 65, no. 3, pp. 401–421, 2016.
- [3] F. Nasirzadeh and P. Nojehdehi, "Dynamic modeling of labor productivity in construction projects," *International Journal of Project Management*, vol. 31, no. 6, pp. 903–911, 2013.
- [4] R. Sonmez and J. E. Rowings, "Construction labor productivity modeling with Neural Networks," *Journal of Construction Engineering and Management*, vol. 124, no. 6, 1998, DOI: 10.1061/(ASCE)0733-9364(1998)124:6(498).
- [5] K. M. El-Gohary and R. F. Aziz, "Factors influencing construction labor productivity in Egypt," *Journal of Management in Engineering*, vol. 30, no. 1, January 2014, DOI: 10.1061/(ASCE)ME.1943-5479.0000168.
- [6] A. M. Jarkas and C. G. Bitar, "Factors affecting construction labor productivity in Kuwait," *Journal of Construction Engineering and Management*, vol. 138, no. 7, pp. 811–820, July 2012, DOI: 10.1061/(ASCE)CO.1943-7862.0000501.
- [7] P. Pornthepkasemsant and S. Charoenpornpattana, "Factor affecting construction labor productivity in Thailand," 2015 International Conference on Industrial Engineering and Operations Management (IEOM), Dubai, United Arab, pp. 1–6, 3–5 March 2015.
- [8] K. N. Hewage and J. Y. Ruwanpura, "Carpentry workers issues and efficiencies related to construction productivity in commercial construction projects in Alberta," *Canadian Journal of Civil Engineering*, vol. 33, pp. 1075–1089, 2006.
- [9] S. Ahmed, M. I. Hoque, M. H. Islam and M. Hossain, "A reality check of status level of worker against skilled worker parameters for Bangladeshi construction industry," *Journal of Civil Engineering and Construction*, vol. 7, no. 3, pp. 132–140, 2018.
- [10] R. Choudhry and B. Zafar, "Effects of skills, motivation, and personality traits on the competency of masons," *International Journal*



- of Sustainable Real Estate and Construction Economics, vol. 1, no. 1, pp.16–30, 2017.
- [11] J. Dai, P. M. Goodrum and W. F. Maloney, “Construction craft workers' perceptions of the factors affecting their productivity,” *Journal of Construction Engineering and Management*, vol. 135, no. 3, 2009, DOI: 10.1061/(ASCE)0733-9364(2009)135:3(217).
- [12] U. C. Gatti, G. C. Migliaccio, S. M. Bogus and S. Schneider, “An exploratory study of the relationship between construction workforce physical strain and task level productivity. *Construction Management and Economics*, vol. 32, no. 6, pp. 548–564, 2014.
- [13] B. G. Hwang, L. Zhu and J. T. T. Ming, “Factors affecting productivity in green building construction projects: The case of SinAGPore,” *Journal of Management in Engineering*, vol. 33, no. 3, 2017, DOI: 10.1061/(ASCE)ME.1943-5479.0000499.
- [14] S. Loganathan, and S. Kalidindi, “Masonry labour construction productivity variation: An Indian case study,” *Proceedings of the First Indian Lean Construction Conference*, Mumbai, India, vol. 1, pp. 175–185, 16-17 February, 2015.
- [15] I. Mahamid, “Contractors perspective toward factors affecting labor productivity in building construction,” *Engineering, Construction and Architectural Management*, vol. 20, no. 5, pp. 446–460, 2013.
- [16] G. J. Sweis, R. J. Sweis, A. A. Abu Hammad and H. R. Thomas, “Factors affecting baseline productivity in masonry construction: A comparative study in the US, UK and Jordan,” *Architectural Science Review*, vol. 51, no. 2, pp. 146–152, 2008.
- [17] Quota No. 1776, Estimating Quota of Construction Projects, Vietnamese Ministry of Construction, Issued date of August 2007. Retrieved 29 February, 2020 at: [http://cucqlxd.gov.vn/media/documents/dinh\\_muc\\_1776.pdf](http://cucqlxd.gov.vn/media/documents/dinh_muc_1776.pdf).

# Chances and Challenges of Vietnam's Garment Industry in the new Trend of Sustainable Development

Tri Tran Quang  
Faculty of Garment Technology and  
Fashion  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tritq@hcmute.edu.vn

Tu Tran  
Faculty of Garment Technology and  
Fashion  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
camtuspkt@hcmute.edu.vn

Alang Tho  
School of Business  
Ho Chi Minh City International  
University  
Ho Chi Minh City, Vietnam  
alangtho@hcmiu.edu.vn

John Burgess  
School of Management  
RMIT University  
Melbourne, Australia  
john.burgess@rmit.edu.au

**Abstract**— To remain internationally competitive, Vietnam's garment industry aims to build upon low production costs with the target of taking available advantages from reasonable pricing, the improvement of product quality, and an increase in labor productivity. However, despite export growth, the added value of the industry is still not as worthy as its potential. This leads to the significance of the study to explore underlying reasons for what hindered the development of Vietnam's garment industry. The study was done following a two-phase study. Phase I was a qualitative exploration through in-depth interviews with eight companies' executives, and Phase II was quantitative research through an online survey from selected leading professors and company executives representing different aspects of Vietnam's garment industry. Six main themes emerging from this study reveal two different viewpoints of Vietnam's garment industry, including both chances and challenges that Vietnam's garment industry will be facing in the current trend of deepening and widening international integration.

**Keywords** — Vietnam, garment industry, chances, challenges, sustainable development

## I. INTRODUCTION

China, Bangladesh, and Vietnam are the three largest textile and garment exporters in the world [1]. The annual export turnover of Vietnam's garment industry reached \$30 billion US dollars in 2019 and became the second largest contributor to the GDP of the country [2]. The growth of Vietnam's garment industry is expected to be continued in the next decade as it plans to gain the target of export turnover at \$60 billion US dollar in 2025. Also, employment in this industry will grow from 2.95 million jobs in 2020 to 3.5 million jobs in 2025 [3].

Despite this considerable growth, previous studies suggest that Vietnam's garment industry lags behind its foreign counterparts such as China and Bangladesh, in terms of prices and labor productivity [4][5][6]. This issue of the garment industry becomes further exacerbated since the impact of

coronavirus shutdown in earlier 2020. Previous studies have focused on the investigation of 'inside' factors of each garment enterprise to find out enablers and hindrances related to the sustainable development of garment industry in Vietnam [4][5][6]. However, in this paper, the argument is that the growth of the garment industry depends on several factors going beyond the organizational context, such as the support of the government, the international market, foreign emerging competitors, inflation, and the domestic market. The research question applied to guide this study is what are the chances and challenges from outside the organizations impacting on the sustainable development of Vietnam's garment industry?

To address the research question, the study was conducted through two phases [7]. The first phase involved interviews with organizations' executives in having their perspectives on these issues. The second phase was a web-based survey conducted with 83 informants who work and research on the garment industry to strengthen the findings from the qualitative research. Due to the lack of SWOT model references in Vietnam's garment industry, this research is considered one of the first inputs for the SWOT model of Vietnam's garment industry succeeding. This study aims to reveal and provide genuine information on the chances and challenges that Vietnam's garment industry is dealing with in an attempt to encourage contributions to understanding its business practices and the global market, leading to sustainable development. The structure of the paper is as follows. Firstly, we present the methodologies applied in this study. Secondly, the key results are outlined. Finally, the discussion and conclusion are presented.

## II. METHODOLOGY

The study was conducted following a sequential mixed method approach, including a two-phase study [7]. The reason for collecting qualitative data initially is that there are deeper and wider judgments, criteria for analyzing because of a

shortage of a guiding theory, and few taxonomies relating to Vietnam's garment industry today.

Phase I was a qualitative exploration through in-depth interviews from eight selected company executives representing different aspects of Vietnam's garment industry [8][9][10]. Data was gathered through interview questions developed by the researcher. The interview questions was structured by a designated set of questions tightly linked with research objectives to find out generalizations in each category. In order to ensure objective and strong data collected, the research sample consisted of a group variety of interviewees representing two different aspects of Vietnam's garment industry such as directors and managers in the manufacturing aspect; and head of school/college and principle in the educational aspect (Table 1).

Phase II was quantitative research through an online survey [11][12]. Based on findings from Phase I, a survey questionnaire was designed to develop further the findings of Phase I. The informants again include those familiar with production/manufacturing; and training/education for the sector. There were more participants as compared to stage 1 and it took in a wider geographical spread of participants. The rationale of choosing those participants and informants was that they were experienced experts and knowledgeable academia in the specific context of Vietnam's garment industry. Due to lack of participants for the field of Vietnam's garment industry, in-depth interviews were used to explore key themes through a small number of participants first before testing those themes through a larger amount of informants via the online survey. The output of phase I was used as the input of the phase II to test the result of the phase I.

TABLE 1. INTERVIEW PARTICIPANT PROFILES

Participants	Title of participant	Name of Company
Participant 1	General Director	Saigon 3 Garment Joint Stock Company Ho Chi Minh City
	Vice President	Vietnam Textile and Apparel Association VITAS
Participant 2	Marketing Manager	Protrade Garment Company
Participant 3	Sales Manager	Sonha limited company
Participant 5	Vice Dean	Faculty of Garment Industry and Fashion
Participant 4	Vice Manager	Phong Phu Corporation
Participant 6	Head of College	Viet Tien Garment Joint Stock Corporation
Participant 7	Head of College	Department of Technical Textile
Participant 8	Principal	Ho Chi Minh City Vinatex Economic Technical College

### III. RESULTS

#### A. The result of Phase I:

##### 1) Category 1: Chances of Vietnam's garment industry

a) *Theme 1:* Having a series of opportunities to expand the export market when Vietnam goes through international trade arrangements: The Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP), and the EU-Vietnam Free Trade Agreement (EVFTA).

Participation in organizations like the CPTPP and the EVFTA will definitely have a positive impact on the development of the garment industry in Vietnam. This obliges Vietnam garment producers to have to change accordingly. From here, it will help Vietnam's garment industry to enhance competitiveness in the fiercely global trend.

Participant 1: Like garment exporters, foreign importers anticipated benefits that the free trade agreements could bring to them. They seek garment exporters who can bring the highest economic efficiency for them. Against, garment exporters also can enjoy market expansion. Many garment exporters are still waiting passively for benefits rather than actively seeking and seizing them.

Participant 2: Since China's wages increased in recent years, customers found other producers who can supply the same quality product and reasonable prices. Therefore, the number of customers moved to Vietnam, even very large customers from China. This led to a great opportunity for Vietnam's garment to expand export markets.

b) *Theme 2:* "Made-in-Vietnam" products have to take advantage of opportunities in the domestic market

After paying more attention to the international market, Vietnam's garment industry seems to realize the potential fertile land of the domestic market. It is taking advantage of the available chances to exploit the domestic market properly and thoroughly.

Participant 2: Previously, when seeing products with the label "Made in Vietnam," in general, Vietnamese people were not interested due to the inferior quality and price of these products. However, it is currently the opposite, particularly for Vietnam's garment products. To give an illustration, for the same products like a shirt, Vietnam product quality is not worse than that of any country about the quality of the seam or the assembly. Therefore, Vietnam's garment products gain trust and have a firm foothold in the minds of domestic consumers.

Participant 4: In recent years, garment products "Made in Vietnam" had a solid foothold on the domestic market. However, to win the total trust in the customers, Vietnam's garment industry has to reduce the product price, diversify in models, and constantly improve product quality.

c) *Theme 3:* To take advantage of the public policy programs that provide special support for the sector.

Vietnam's garment industry plays a key role in solving the biggest problem of employment of the huge workforce, contributing to social stability and development. It is one of the reasons why there are a lot of emerging sectors. However, the garment industry is still receiving special concern from the government via its incentive investment and development.

Participant 4: The garment industry in Vietnam is considered a priority sector to be encouraged by incentive investments and development because every year, it has brought the biggest export value to Vietnam.

Participant 2: According to the Prime Minister policy about the development of Vietnam's garment industry until 2020, the garment sector orients up to 3 million workers in 2020. This shows the great concern of the Vietnam government about the importance of Vietnam's garment industry in bringing practical effect to the development of the Vietnam economy.

2) *Category 2: Challenges of Vietnam's garment industry*

a) *Theme 4: Fierce competition from China, India, and Bangladesh*

Garment industry exporters like China, India, Bangladesh had invested and developed for a long time but still have been investing in this sector. They are much better than Vietnam in research, training, human resource development, advanced machinery and equipment and capability of meeting orders required the high technical quality due to the more synchronous investment comparing to those of Vietnam.

Participant 1: Garment enterprises are coping with the competition of big garment exporting countries such as India, Bangladesh. Due to the global recession, these countries are offering competitive prices to attract customers. Recently, the tendency of moving orders from China to neighboring countries is reducing because China's economy is struggling, and China garment enterprises must retain orders for production. With the drawback of being passive in material supply sources, it is clear that Vietnam's garment enterprises will encounter challenges.

Participant 4: Like Vietnam, big garment industry exporters such as India, China, and Bangladesh have their advantages namely the population density and the cheap labor cost, but they also have many favorable factors to expand the production in this sector such as experience, machinery, material supply sources, et. Therefore, they dominate the majority of the garment industry market segment is not surprising.

Participant 8: The competitiveness of powerful exporters such as China, India. Especially China is the most formidable opponents because of three main reasons:

- First, China labor workforce is not inferior to Vietnam but better disciplined.
- Second, it has adequate raw material and machinery sources.
- Third, it has better macroeconomic management

b) *Theme 5: Potential emerging competitors like Cambodia, Laos, Myanmar*

Countries with an emerging garment Industry with the advantage of cheaper labor costs such as Cambodia, Myanmar, Laos (with wages of around 100 USD per month) are also new challenges of Vietnam's Garment Industry.

Participant 3: Before ordering in a certain garment company, customers had the research and comparison among companies from many different countries. Especially Cambodia has currently emerged as a direct competitor to Vietnam. Because labor costs in Cambodia

are very cheap, just over \$ 100 per month and Cambodia is still a free market, so there is no tax to import raw materials.

Participant 1: The EU is moving away to import orders from Vietnam to Cambodia, Laos, to avoid import tax rates of 10% because these countries are entitled to benefit from the most-favored-nation (MFN) tax rate of 0% from EU imports for developing countries. Myanmar is also the free market, with simple administrative procedures and very low-price labor cost. Moreover, because Myanmar has normalized diplomatic relationship with the United States, there are more favorable incentives on the import and export tax

c) *Theme 6: Increasing costs*

The high inflation rate in Vietnam's current situation is the global challenge in general and in Vietnam's garment industry in particular. Thus, to reduce the inflation rate at the lowest level, Vietnam should have concentrated efforts to prepare and cope effectively with any difficult situation happening to the economy in the context of the global recession.

Participant 1: Many enterprises are under great pressure due to inflation, leading to difficulties in the production process. In the profit of enterprises, wages accounted for 65%. And the new policy on adjustment of wages to stabilize the lives of workers from the beginning of October 2011 is creating more pressure on enterprises, especially those large enterprises having thousands of workers.

Participant 2: Because the garment industry serves the basic needs of human life like the dressing, fluctuations in price caused by inflation affected consumer needs negatively. And Vietnam's inflation is approximately 15% per year. This means next year, for workers to be able to live like today, the company must pay an additional 15% while our enterprise cannot increase in the price of orders from customers. This creates a huge pressure on not only enterprises in garment industry sectors but also other sectors of Vietnam.

B. *The result of Phase II:*

Table 2 presents three chances for Vietnam's garment industry including "Having a series of opportunities to expand the export market when Vietnam goes through international trade arrangements: The CPTPP and the EVFTA" (average value at 72.29), "Made-in-Vietnam products have to take advantage of opportunities in the domestic market" (average value at 56.63), and "To take advantage of the public policy programs that provide special support for the sector" (average value at 60.24). For the challenges (Table 3), there were three factors including "Fierce competition from China, India, and Bangladesh" (average value at 77.11), "Potential emerging competitors like Cambodia, Laos, Myanmar" (average value at 55.42), and "Increasing costs" (average value at 68.67). All these findings addressed the statistical significance (P-Value at 0.01).

TABLE 2. CHANCES FOR VIETNAM'S GARMENT INDUSTRY

Chances	0	1	2	3	4	Average Value
1. Having a series of opportunities to expand the export market when Vietnam goes through international trade arrangements: The CPTPP, and the EVFTA.	1	2	20	24	36	72.29
2. Made-in-Vietnam products have to take advantage of opportunities in the domestic market.	0	10	26	30	17	56.63
3. To take advantage of the public policy programs that provide special support for the sector.	0	11	22	36	14	60.24

(Using a scale of 0 = Not at all important to 4 = Very important)

TABLE 3. CHALLENGES FOR VIETNAM'S GARMENT INDUSTRY

Challenges	0	1	2	3	4	Average Value
4. Fierce competition from China, India, and Bangladesh	0	2	17	16	48	77.11
5. Potential emerging competitors like Cambodia, Laos, Myanmar.	1	14	22	17	29	55.42
6. Increasing costs	0	2	24	26	31	68.67

(Using a scale of 0 = Not at all important to 4 = Very important)

#### IV. DISCUSSION AND CONCLUSIONS

The purpose of this study is to explore the challenges and chances that impact the sustainable growth of Vietnam's garment industry. We found that the Vietnam's garment industry had several chances to grow, but they also faced with challenges.

Fierce competition from China, India, and Bangladesh was judged the enormous challenge of Vietnam's garment industry, with up to 77% of respondents. The competitiveness of powerful exporters such as China, India, and Bangladesh are a big challenge. Especially China is the most formidable opponents of Vietnam's garment industry. However, Vietnam's garment industry has a series of opportunities to expand the export market when Vietnam goes through international trade arrangements: the CPTPP and the EVFTA. This is the biggest chance of Vietnam's garment industry in the next coming year that is approbated by 72% of respondents. Participating organizations such as the CPTPP and the EVFTA will definitely have a positive impact on the development of the textile industry in Vietnam. This obliges Vietnam garment producers to change accordingly. From that point, it will help Vietnam's garment Industry to enhance their

competitiveness in the fiercely global trend. The most important thing now is that Vietnam should actively exploit the gaps in market segments from its intermediate level or less. At the same time, Vietnam needs to reduce waste in the production process and stabilize material supply sources to be able to compete with these competitors.

Potential emerging competitors like Cambodia, Laos, and Myanmar in the future challenge with the rate of 55% of respondents. EU begins moving away from Vietnam to Cambodia and Laos for import orders to avoid import tax rates of 10% because these countries are entitled to benefit from the most-favored-nation (MFN) tax rate of 0% for EU imports for developing countries which will affect to the development of garment industry in Vietnam. On the other hand, to turn challenges into opportunities, "Made-in-Vietnam" products have to take advantage of opportunities in the domestic market" is evaluated by respondents at the lowest rate among chances only 57%. In recent years, T&C products "Made in Vietnam" is gradually retrieving the foothold on the domestic market from foreign competitors. However, to achieve this thoroughly, Vietnam's garment industry needs to make distinctions for its products like the quality of its product and the prestige with customers to form a good basis for the competition. Also, it has to continuously reduce the product price, diversify in models suiting the taste of domestic customers, and constantly improve product quality to win the total trust in the customers.

Increasing cost is considered one of the challenges for the garment industry in the current time, receiving the interest of 69% of respondents. The high level of inflation in the current situation is also the challenge for Vietnam in general and garment industry in particular. Since the garment Industry serves the basic needs of human life like the dressing, the variation in price caused by inflation negatively affected this sector. On the other hand, Vietnam's garment industry is not only the great benefit brought from exports but also the most labour intensive industry, contributing to solving the biggest problem of employing the huge labor workforce of Vietnam and social welfare. So, 60% of respondents saw opportunities for the sector to take advantage of incentive investment and development funding provided by the government. Hence, Vietnam's garment industry needs to effectively and comprehensively exploit the economic policies to enhance the added value and profits on each garment product and improve living standards for workers in this sector.

#### IV. RESEARCH LIMITATIONS

The interviews involved a limited number of participants, as did the survey. The interviews were confined to the region of South Vietnam. Other stakeholders such as government and garment buyers were not included. The research was conducted at one point of time

#### V. CONTRIBUTION OF THE STUDY

The findings of this study will support the industry and policymakers to identify and address opportunities and challenges facing with the sector in the global market.

#### REFERENCES

- [1] Sai Gon Giai Phong Online, "Det may Viet Nam phai dung top dau the gioi (Vietnam's Textile and Garment Industries must be the top group of the world)," Accessed online at <https://www.sggp.org.vn/det-may->



- [viet-nam-phai-dung-top-dau-cua-the-gioi-634871.html](http://viet-nam-phai-dung-top-dau-cua-the-gioi-634871.html), Accessed date 03/05/2020.
- [2] Doanh Nhan Viet, "Dua Viet Nam tro thanh top dau the gioi ve det may (Promote Vietnam's Textile and Garment Industries being the top group of the world)," Accessed online at <https://doanhnhnviet.news/chinh-tri-thoi-su/dua-viet-nam-tro-thanh-top-dau-the-gioi-ve-det-may-6732.html>, Accessed date 03/05/2020.
- [3] Dau Tu Chung Khoan, "Doanh nghiep dey may huong toi muc tieu phat trien ben vung (Vietnam's Textile and Garment Enterprises lead to the sustainable development)," Accessed online at <https://tinnhanhchungkhoan.vn/thuong-truong/doanh-nghiep-det-may-huong-toi-muc-tieu-phat-trien-ben-vung-311270.html>, Accessed date 03/05/2020.
- [4] H. Hill, "Vietnam textile and garment industry: notable achievements, future challenges," Appendix II of the Industrial competitiveness review, 1998.
- [5] Nadvi, K., Thoburn, J. T., Bui Tat, T., Nguyen Thi Thanh, H., Nguyen Thi, H., Dao Hong, L., & Blanco De Armas, E, "Vietnam in the global garment and textile value chain: Impacts on firms and workers," *Journal of International Development*, 2004, 16(1), pp. 111–123.
- [6] Vixathep, S. and Matsunaga, N., "Firm performance in a transitional economy: a case study of Vietnam's garment industry," *Journal of the Asia Pacific Economy*, 2012, 17(1), pp.74-93.
- [7] W. C. John, "Research Design: Qualitative, Quantitative, and Mixed Methods Approaches," Third Edition, SAGE Publications Inc 2009.
- [8] Flick, U, "An introduction to qualitative research," London, United Kingdom: Sage Publications Ltd, 2014.
- [9] Bryman, A., & Bell, E, "Business research methods," New York, NY, United States of America: Oxford University Press, 2015.
- [10] Miles, M. B., & Huberman, A. M., "Qualitative data analysis: An expanded sourcebook (2nd ed.)," Thousand Oaks: Sage Publications, 1994.
- [11] Qualtrics, "Qualitative vs Quantitative Research: What is it and when should you use it?" Accessed online at <https://www.qualtrics.com/blog/qualitative-research/>, Accessed date 06.05.2020.
- [12] Qualtrics, "More students and faculty rely on Qualtrics than any other research platform," Accessed online <https://www.qualtrics.com/education/student-and-faculty-research/>, Accessed date 06.05.2020.

# Extraction of Pectin from *Passiflora edulis* by Aqueous Two-Phase System

Nga Thi Vo

Faculty of Chemical and Food  
Technology

Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
ngavt@hcmute.edu.vn

Thi Hao Cao

Faculty of Chemical and Food  
Technology

Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
caohao0102@gmail.com

Minh Hao Hoang

Faculty of Chemical and Food  
Technology

Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
haohm@hcmute.edu.vn

**Abstract**—Pectin, a thickening additive widely used in food, is present in a large amount in the passion fruit peel. The extraction of pectin for the industrial applications would give the economic efficiencies from waste peel and decrease the risk of pollution. However, traditional extraction methods are energy-consuming procedures that affect the purpose of scrap utilization. In the current work, we reported a method to overcome drawbacks by using an aqueous two-phase-extraction. The appropriate conditions for maximization of pectin yield were the use of a system of potassium carbonate (19.93%, w/w); ethanol (28.77%, w/w); water (51.30%, w/w); passion peel extract (0.0200 g/mL) at room temperature. The application of these parameters for the extraction procedure delivered pectin in a 70.91% yield.

**Keywords**—Aqueous two-phase extraction (ATPE), *Passiflora edulis*, pectin, ethanol/potassium carbonate,

## I. INTRODUCTION

The tropical climate produces luxuriant fruits, and the passion fruit (*Passiflora edulis*) is one of the succulent fruits and widely consumed in Southeast Asia. The fruit can be used directly or made into drinks or incorporated into the other products in the food industry [1]. The passion fruit peel is known as a rich resource of pectin. The main applications for pectin are as a gelling agent, thickening agent, and stabilizer in the food processing industry [2]. Considering the benefits of pectin to the food industry, the extraction procedure of pectin, as a by-product of the passion fruit processing is researched to increase the economic efficiency and reduce the risk of pollution.

Pectin, known as polysaccharide, comprises of galacturonic acid units, which are esterified partially with methanol. Pectin extraction is a multiple-stage in which the hydrolysis and extraction are dependent on various factors, mainly temperature, pH, and time. Pectin is commonly extracted by dilute acid solutions such as hydrochloric, citric, and nitric acids. The resulting hydrolysis product precipitates in ethanol. The pectin extraction was obtained in a 70% yield using citric acid, while the ones were 38% and 26% in nitric acid and hydrochloric acid, respectively.[3]

The aqueous two-phase system (ATPS) was accidentally explored by Martinus Willem Beijerinck, in 1896, during the experiment of mixing an aqueous starch solution with gelatin. However, until the 1950s the application of an aqueous two-phase system (ATPS) became realistic by Per-Åke Albertsson's experiments. The interesting results introduced a new strategy for the extraction of biomolecules, the aqueous two-phase extraction (ATPE) method. The ATPE was a liquid-liquid fractionation technique, in which water was a

primary component of both phases. There were four popular biphasic systems, such as polymer/polymer, polymer/salt, ionic surfactant/ionic or non-ionic surfactant, and ionic liquid/short-chain alcohol. The success of this method up to now is laudable, and it has widely applied to separate and purify various biological products such as cells, nucleic acid, enzymes, proteins [4,5].

The ATPE has been demonstrated as a friendly environment extraction method because it can reduce energy and organic solvent consumption. On the other hand, this method is flexible and adaptable to apply in various fields and easy to scale up. Besides these benefits, the method still exists some disadvantages that the ATPE technique is complicated to find out the optimal condition for biphasic separation and difficult to predict. Therefore, it is necessary to survey each specific case before implementation.

The previous pectin extraction researches using acidic solution reported that the yield of pectin was obtained at 65 – 90°C and pH 1.5–4, with extraction time 1-2 hours, and the pectin precipitation in ethanol needed about 16-24 hours [2,3,6,7]. Clearly, the acidic extraction method is a time-consuming procedure. This work aimed to extract pectin from passion fruit peel using the ATPS of ethanol/salt system, in which the color contaminates would go to the upper phase (ethanol-rich phase), while pectin would precipitate and present at the interface. Therefore, using the ATPE method would decrease the pectin extraction time and energy costs.

## II. MATERIAL AND METHOD

### A. Material and reagents

#### 1) Material and sample preparation

Preparation of crude passion fruit peel powder: The raw powder was prepared as follows: The collected fruits were washed and cut into two portions. The fruit flesh inside the fruits was removed to obtain the peels. The peels were divided into small species and dried in a drying oven (Binder drying oven FD 56, Germany) at 55°C until a constant weight was achieved. The dried samples were ground using an electric miller (Yamafuji DE-500, Japan) and passed through 120 mesh to obtain a fine powder. The ground powder was packaged in polyethylene bags and stored at -18°C in a fridge freezer (Arctiko LTF 425, Denmark) for further experiments.

Preparation of passion fruit peel suspension: The suspensions were prepared for the experiments by the following method: the dried passion fruit peel powder was mixed with distilled water, then vortexed thoroughly to make suspensions with certainly expected concentrations for each experiment.

## 2) Chemicals

All chemicals and solvents used in all experiments were of analytical grade and purchased from Xilong Scientific, China.

### B. Methods

#### 1) ATPS extraction

##### a) Preparation of ATPS:

An aqueous two-phase system was implemented as follows: a certain mass of salt was dissolved in a given amount of distilled water, then a pre-determined volume of absolute ethanol was added. The obtained mixture was mixed and held until reaching equilibration for phase separation.

##### b) Construction of phase diagram:

The phase diagram of ethanol and potassium carbonate was constructed by a turbidity titration method at ambient temperature described by Bensch M. with a slight adjustment.[8] First, a pre-determined mass of salt was dissolved in a given amount of water in a conical flask (at the saturated concentration of each salt). Next, the conical flask of salt solution was put on an analytical balance (Sartorius TE214S, Germany) and tared. Then, absolute ethanol was added in a dropwise manner from a burette to the salt solution. The weight of ethanol was recorded when the first drop of ethanol has just made the salt solution turbid spontaneously. A few drops of distilled water were added to disappear the turbidity, and further absolute ethanol was dropped to reach the next turbid point. The above procedure was repeated until a large amount of ethanol cannot precipitate the salt solution anymore. Both the amounts of ethanol and water added at the different turbid points were recorded exactly. The percentage concentrations of the compositions including ethanol, salt, and water in the mixtures at the turbid points were calculated. The phase diagram was achieved as a binodal curve by plotting the above turbid points.

##### c) Preparation of pectin with ATPE

A pre-determined weight of potassium carbonate was dissolved in certain water. A known amount of anhydrous ethanol and a calculated suspension of the passion peel powder were added to the salt. Next, the mixture was stirred well by a magnetic stirrer (IKA RET control-visc, Germany) at room temperature to dissolve the salt completely. Then the whole system was held for about thirty minutes until the two phases spontaneously separated. The bottom phase is a pectin contained-suspension, and the top phase was an ethanol-rich solution containing undesired color material and other impurities.

The bottom phase and the middle layer were filtrated using a vacuum pump filter (KNF N022 AN.18, Germany) to obtain crude pectin as the residues. The filtrated pectin was soaked in ethanol 80% for 4 hours and then filtrated to discard the remaining salt and unwanted color, three times repeatedly. The resulted pectin was dried in an air-forced convection drying oven at 60°C (UF110, Memmert) until acquiring the constant weight.

#### 2) Analytical procedure

##### a) Determination of pectin purity (P)

The pectin purity of the resulted product was determined by the calcium pectate method according to the previous study.[9] Firstly, 0.15 g dried resulted pectin was added into a conical flask containing 100 mL distilled water, then stirred using the magnetic stirrer. Later on, the suspension was mixed

with 100 mL NaOH 0.1 N stirred overnight for full saponification of the pectin solution. Neutralization of the saponified product using 50 mL acetic acid 1N followed by calcium pectate precipitation by treatment of 50 mL CaCl<sub>2</sub> 2 N solution and keep stable for one hour. The resulting solution was boiled for 5 minutes, then the precipitated pectin was filtrated using a Whatman No-1 filter paper (dried until a constant mass achieved and recorded the weight) and washed with boiling water several times. The precipitated pectin holding on the filter paper was dried in the convection oven at 60°C until reaching the constant weight. The mass of the precipitated calcium pectate was calculated using the mass of the whole filter paper with residual minus the weight of the filter paper. The purity of pectin in percentage, P (%), was determined as the following equation:

$$P = \frac{m \times 0.92 \times 100}{B} (\%)$$

When m was the mass (g) of the precipitated calcium pectate, 0.92 was a transformation factor for changing from calcium pectate to pectin [9], and B was the mass of pectin (g) for saponification, which was calculated in this experiment as follows:

$$B = \frac{n \times V_2}{V_1} (g)$$

When n was the mass of pectin to prepare 100 mL suspension, n = 0.15 g; V<sub>1</sub> was the volume of pectin suspension, V<sub>1</sub> = 100 mL; and V<sub>2</sub> was the volume of pectin suspension for saponification, V<sub>2</sub> = 20 mL.

##### b) Estimation of passion pectin recovery yield (Y)

The recovery yield of pectin was calculated using the following equation:

$$Y = \frac{m \times 100}{M \times V} \times P (\%)$$

Where Y was pectin recovery yield (%), m was the mass (g) of pectin obtained after dried, M was the concentration (g/mL) of a suspension sample, V was the volume (mL) of a suspension sample, P was the purity of pectin in the resulted pectin (%).

##### c) Determination of degree of esterification (DE)

The degree of esterification (DE) was an important parameter relating to the structural characteristic of pectin and defined to be a percentage of the esterified galacturonic acid units of the total number of galacturonic acid units in the molecule. The DE of the pectin product was determined using the titrimetric method described in Food Chemical Codex with minor modification.[10] A dried pectin sample of 0.2 g was placed in a conical flask with several drops of ethanol, and then 20 mL of distilled water. The mixture was stirred continuously at 40°C until the pectin dissolved totally. The sample was titrated with 0.1 N sodium hydroxide and three drops of phenolphthalein as an indicator. The initial titration volume of sodium hydroxide (I<sub>t</sub>, mL) was recorded when the pink color appeared freshly. Next, 10 mL of 0.1 N sodium hydroxide was added to the titration flask then closed with a stopper, shaken thoroughly, and left at the ambient temperature for 2 hours until the pectin saponification was completed. Later on, the sample was added 10 mL of 0.1 N hydrochloric acid to neutralize the rest of the sodium hydroxide and sharked until the pink color disappeared. Finally, the sample was added three drops of phenolphthalein

and titrated further with 0.1 N sodium hydroxide. Once the pink color appeared persistently after vigorous shaking, record the final titration volume of sodium hydroxide solution ( $F_t$ , mL). The DE was calculated from the following equation:

$$\%DE = \frac{F_t}{(I_t + F_t)} \times 100$$

When DE was the degree of esterification of pectin,  $I_t$  and  $F_t$  were the initial and the final titration volumes of 0.1 N sodium hydroxide, respectively.

All the analyses were performed in triplicate.

### III. RESULTS AND DISCUSSION

#### A. Phase diagram

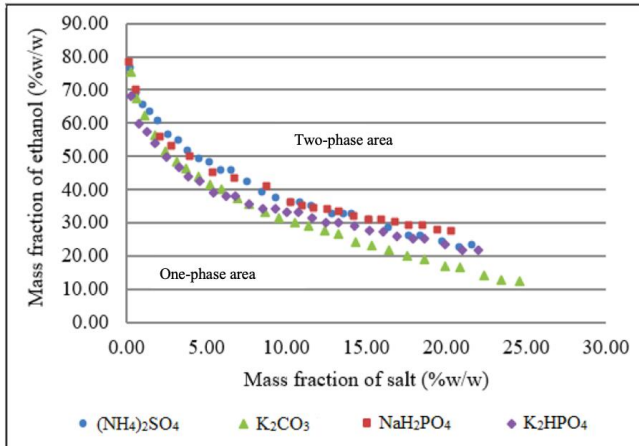


Fig. 1: The phase diagrams of four ATPSs

The purpose of constructing a phase diagram of ATPS was an obvious approach to select the appropriate ratio of phase compositions in each ATPS. The phase diagrams of four different ATPSs including ethanol/ $K_2CO_3$ , ethanol/ $NaH_2PO_4 \cdot 2H_2O$ , ethanol/ $(NH_4)_2SO_4$ , and ethanol/ $K_2HPO_4 \cdot 3H_2O$  were investigated and shown in Fig.1 as binodal curves. Each binodal curve delineated the composition area of each ATPS into two zones, of which the above zone existed two distinguished phases, and the below zone was mono-phasic. In the two-phase area, the top phase was an ethanol-rich aqueous phase and the bottom phase was a salt-rich aqueous phase. The ratio of phase compositions in each ATPS in the following experiments was chosen from the above area of the binodal curve. In practice, increasing the concentration of ethanol until the extreme phase-forming points resulted salting-out.

#### B. Selection of the composition of the ATPS

To approach the optimal ATPE condition for the best enrichment capacity of pectin extraction, four extraction parameters were investigated, namely, phase compositions including phase forming-salt composition and phase forming-ethanol composition, extraction time, extraction temperature, and concentration of the sample's suspension.

##### 1) Selection of phase forming-salt:

Based on the phase diagram, the detailed experiments of four ATPSs with the sample were carried out to select the optimized salt for the highest yield of pectin extraction. Each experiment was performed with the composition of 3.0 mL of distilled water, 3.0 mL of absolute ethanol, 1.0 g of passion fruit peel suspension (0.25 g/mL), and a certain amount of salt (detailed in Table 1). The ATPS parameters of each

experiment were chosen in the two-phase area (Fig. 1). The result from the investigation of four different ATPSs was displayed in Table 1. It was seen that the ATPS of ethanol/ $K_2CO_3$  owned the maximum yield of pectin extraction (57.73%) which indicated that the ethanol/ $K_2CO_3$  system possessed the best enhancement capacity for pectin extraction from the bottom phase. Therefore, the ATPS of ethanol/ $K_2CO_3$  was chosen for further investigated experiments.

TABLE I. EXTRACTION OF PECTIN USING FOUR DIFFERENT ATPSs OF ETHANOL/SALTS

Ethanol/salt system	Salt added (g)	Mass of pectin (g)	Yield (Y, %)
Ethanol/ $(NH_4)_2SO_4$	0.6	$0.0056 \pm 0.0003$	22.53
	0.7	$0.0068 \pm 0.0004$	27.33
	0.8	$0.0072 \pm 0.0001$	28.93
	0.9	$0.0080 \pm 0.0004$	32.00
	1.0	$0.0068 \pm 0.0004$	27.20
Ethanol/ $NaH_2PO_4$	1.8	$0.0092 \pm 0.0003$	36.67
	2.0	$0.0100 \pm 0.0004$	40.00
	2.2	$0.0106 \pm 0.0003$	42.53
	2.4	$0.0098 \pm 0.0004$	39.07
	2.6	$0.0089 \pm 0.0003$	35.47
Ethanol/ $K_2HPO_4$	1.8	$0.0076 \pm 0.0003$	30.27
	2.0	$0.0081 \pm 0.0003$	32.27
	2.2	$0.0091 \pm 0.0003$	36.27
	2.4	$0.0099 \pm 0.0006$	39.47
	2.6	$0.0108 \pm 0.0005$	43.07
Ethanol/ $K_2CO_3$	2.0	$0.0111 \pm 0.0006$	44.40
	2.2	$0.0126 \pm 0.0001$	50.53
	2.4	$0.0133 \pm 0.0004$	53.20
	2.6	$0.0144 \pm 0.0002$	57.73
	2.8	$0.0137 \pm 0.0004$	54.80

Each experiment was performed with the composition of 3.0 mL of distilled water, 3.0 mL of absolute ethanol, 1.0 g of passion fruit peel suspension (0.25 g/mL), and a certain amount of salt (detailed in Table 1).

##### 2) Effect of phase compositions:

###### a) Effect of phase forming-salt composition:

Applying the selected salt,  $K_2CO_3$ , the influences of phase forming-salt concentration on the yield of pectin in ATPS were studied (Fig. 2).

Experimentally, when applying the passion peel sample to the ATPSs of  $K_2CO_3$  and ethanol, the extreme phase forming points of the salt were 16.46 - 24.46% (w/w), and of the ethanol 21.63 - 35.73% (w/w). If the ethanol concentration was less than 21.63%, no two-aqueous phase could be formed, whereas when the concentration was above 35.73%,  $K_2CO_3$  would precipitate in the system. Therefore, the experiments for the effect of phase forming-salt were designed with the concentration of ethanol as 28% (w/w).

The designed experiments for the effect of phase forming-salt on the recovery of pectin were prepared with the composition as following: the ATPSs including 15 mL of distilled water, 28% (w/w) absolute ethanol, constantly for each experiment, and the pre-determined salt (w/w). The 5 mL of passion fruit suspension at a concentration of 0.0250 g/mL was added to the prepared ATPS. In these experiments, the ethanol concentration was fixed at 28%, while the  $K_2CO_3$  concentrations increased from (w/w) 16.46% to 24.46%. The above-limited range of salt was less than 16.46% of potassium carbonate, the mixture of ATPS and passion peel suspension

could not form two phases, whereas the above 24.46% of salt, the potassium carbonate would precipitate in the system.

With the increase of  $K_2CO_3$  concentration in ATPS, the yield of pectin was enhanced and reached a 50.96% highest yield with the purity ( $P = 86.07\%$ ) of resulted pectin at a concentration of 19.93%  $K_2CO_3$ . The result could be attributable to the affinity between water and salt. Salt has attracted water from the top phase to the bottom one. The increase of the bottom water-rich phase volume would enhance the extraction of pectin which is dissolvable in water. However, when the amount of salt increased greatly, there is a competition between salt and pectin to coexist in the bottom aqueous phase. This resulted in a decrease in the level of pectin in the lower phase. Therefore, the mass fraction of  $K_2CO_3$  was used for the following investigated experiments was 19.93%.

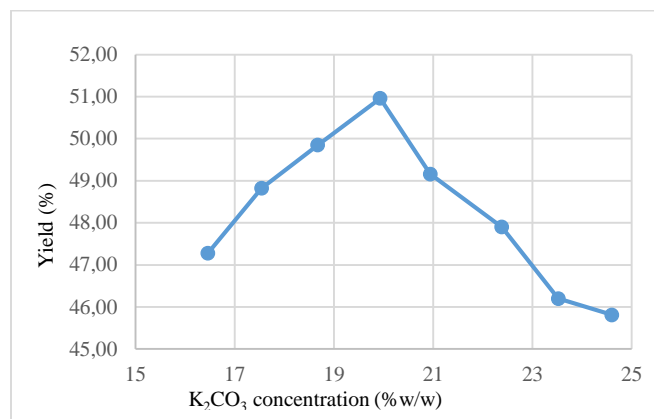


Fig. 2. Effect of  $K_2CO_3$  concentration on the yield of pectin

Each experiment was performed with the composition of 15 mL of distilled water, 28% (w/w) absolute ethanol, the pre-determined salt from 16.46% to 24.46% (w/w), 5 mL of passion fruit suspension at a concentration of 0.0250 g/mL.

#### b) Effect of phase forming-ethanol composition:

The experiments designed for the effect of phase forming-ethanol on the yield of pectin were similar to the ones to investigate the effect of phase forming-salt. The  $K_2CO_3$  concentration was fixed at 19.93%, while the ethanol concentrations were varied in a range from 21.63 to 35.73% (w/w). The limited range of ethanol was determined practically because the two phases could not be separated if the ethanol concentration was lower than 21.63%, whereas the above 35.73% ethanol made the potassium carbonate would salt out the system. The influences of phase forming-salt concentration on the yield of pectin in ATPS were investigated and displayed in Fig. 3.

The results from the experiments showed that with the increase of ethanol concentration from 21.63% to 35.73%, the yield of pectin increased and reach a maximum yield of pectin (50.15%) with the purity of pectin ( $P = 92.41\%$ ) when the ethanol content was 28.77%. When the level of ethanol increased, pectin is transferred to the bottom phase because it was precipitated in the high ethanol concentration. However, when the concentration of ethanol is too high, the balance between two phases was broken because of the salting-out which reduced the bottom phase volume leading to the reducing yield of pectin. Therefore, the phase forming compositions were established as 28.77% (w/w) ethanol and 19.93%  $K_2CO_3$  for the further investigated experiments.

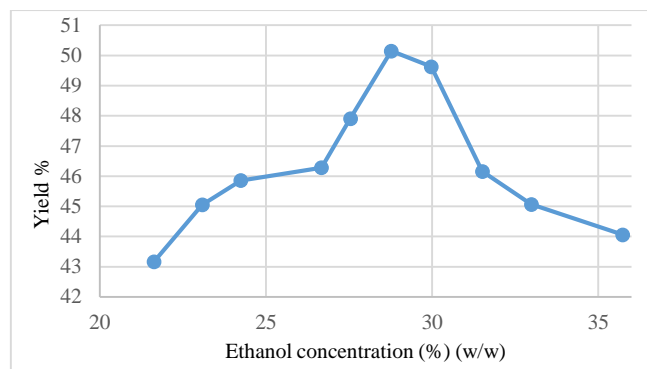


Fig. 3. Effect of ethanol concentration on the yield of pectin

Each experiment was performed with the composition of 15 mL of distilled water, 19.93% (w/w)  $K_2CO_3$ , the pre-determined ethanol from 21.63 to 35.73% (w/w), 5 mL of passion fruit suspension at a concentration of 0.0250 g/mL.

#### 3) Effects of extraction time

Extraction time was an essential parameter affecting the yield of pectin in ATPE. The investigation was carried out with 7 periods of stirring times including 1, 2, 5, 10, 15, 20, and 30 minutes. The experiments were performed similarly to the ones for the effect of phase forming-salt with the previously investigated compositions for ATPS as following 15 mL distilled water, 28.77% (w/w) ethanol, 19.93% (w/w)  $K_2CO_3$ , and 5 mL passion peel suspension 0.0250 g/mL. The experiment results were shown in Fig. 4.

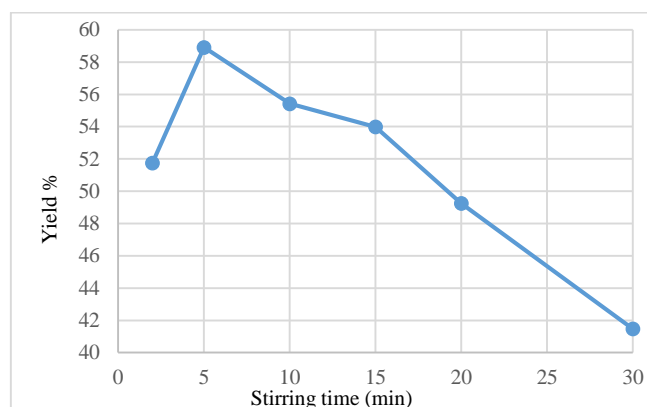


Fig. 4. Effect of extraction time on the yield of pectin

Each experiment was performed with the composition of 15 mL distilled water, 28.77% (w/w) ethanol, 19.93% (w/w)  $K_2CO_3$  and 5 mL passion peel suspension 0.0250 g/mL.

It could be seen that the optimal extraction time was 5 minutes reaching the highest yield of pectin as 58.91% and the purity  $P = 91.59\%$ . The stirring time less than one minute could not be available for phase forming while a longer stirring than 5 minutes, the yield of pectin decreased. The soluble characteristics of pectin were solubility in water and precipitation in ethanol. In the ATPE process applied with a passion peel sample, pectin was precipitated by ethanol and separated as suspension. When the more time stirring, the more pectin was soluble in water and entered into the bottom phase leading to the decreasing yield of pectin. Therefore, the extraction time was selected for the next experiment was 5 minutes.

#### 4) Effects of sample's suspension concentration

In the previous experiments, the concentration of passion fruit peel suspension was 0.0250 g/mL. In current



experiments, the passion peel concentration was scanned in a range from 0.0143 to 0.0500 g/mL to investigate the effect of suspension concentration on the yield of pectin. The ATPS comprised 28.77% (w/w) ethanol and 19.93% (w/w)  $K_2CO_3$  and 7.5 mL passion peel suspension with different concentrations was added. The practical results were displayed in Fig. 5.

As the concentration of passion fruit peel suspension increased, the extraction efficiency increased and peaked at the sample concentration of 0.0200 g/mL with the maximum yield of 70.91% and the purity of pectin P = 92.41%. When the sample concentration increased by more than 0.0200 g/mL, the yield decreased. This was due to the extraction capacity of the system. Initially, the system was still responding to the increase of sample concentrations, but when the sample concentration exceeded the limited capacity of the system, the yield decreased. Therefore, the optimal concentration of passion fruit peel suspension was determined as 0.0200 g/mL.

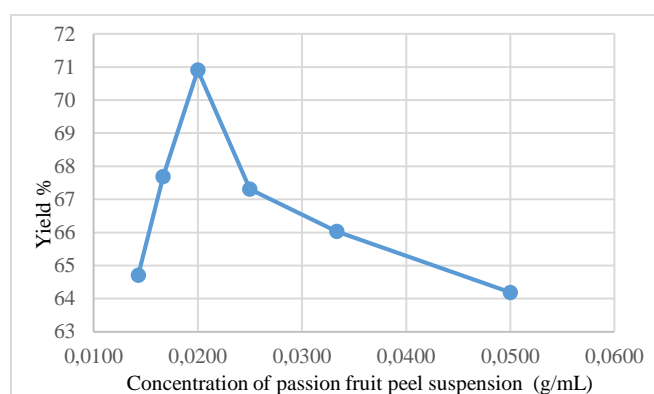


Fig. 5. Effect of concentration of suspension on the yield of pectin

Each experiment was performed with the composition of 15 mL distilled water, 28.77% (w/w) ethanol, and 19.93% (w/w)  $K_2CO_3$  and 7.5 mL passion peel suspension with different concentrations.

### C. Characteristic of the resulted pectin

The pectin product obtained was slightly yellowish, tasteless, odorless.

The degree of esterification (DE) of the pectin product was determined as 74.67% belonging to the high methoxy (HM) pectin type which was used in high sugar products. The result was similar to the one from other studies indicating that the DE of pectin extracted from passion fruit peel by an acidic extraction method using citric acid was 78.59% [2].

The pectin yields by ATPE were compared with the ones by the acidic extraction methods using passion fruit peel or others as starting materials. The previous studies have resulted that citric acid was the most effective for pectin extraction. Erika Kliemann approached the optimum factors to yield the maximum pectin of 70% when using acidic extraction of passion fruit peel at 80°C [3]; Calliari obtained the yield of

pectin of 77% when extracting citrus peel at 100°C [6]; Virk achieved a 78% yield of pectin from apple pulp [7]. Upon the above information, the application of ATPE on pectin extraction gave similar efficiencies with the acidic extraction method. The advantage of the ATPE process is energy reduction because the extraction was processed at ambient temperature.

## IV. CONCLUSION

The whole investigation of parameters effecting on the yield of pectin led to the optimal extraction conditions as follows: ATPS composition (w/w) composed of ethanol (28.77%),  $K_2CO_3$  (19.93%), passion fruit peel suspension concentration (0.0200 g/mL), extraction time (5 minutes). The yield of resulted pectin was 70.91% with a purity of 92.41%, and the degree of esterification was 74.67%. ATPE could become a promising extraction method that is applicable to the food industry due to its advantages as a simple method to apply and low energy cost process.

## ACKNOWLEDGMENT

The authors are thankful to HCMC University of Technology and Education for providing facilities to complete our study.

## REFERENCES

- [1] Julia F. Morton, "Passionfruit" in Fruits of warm climates, New Crop Resource Online Program, Purdue University, 1987, pp. 320–328.
- [2] Eloisa Rovaris Pinheiro et al., "Optimization of extraction of high-ester pectin from passion fruit peel (*Passiflora edulis flavicarpa*) with citric acid by using response surface methodology", *Bioresource Technology*, Vol 99, 2006, pp. 5561 – 5566.
- [3] Erika Kliemann et al., "Optimisation of pectin acid extraction from passion fruit peel (*Passiflora edulis flavicarpa*) using response surface methodology", *International Journal of Food Science and Technology*, vol 44, 2008, pp. 476 – 483.
- [4] Mujahid Iqbal et al., "Aqueous two-phase system (ATPS): An overview and advances in its applications", *Biological Procedures Online*. Vol 18, 2016. [DOI 10.1186/s12575-016-0048-8].
- [5] João Vitor Dutra Molino et al., "Different types of aqueous two-phase systems for biomolecule and bioparticle extraction and purification", *Biotechnology Progress*, 2013, pp.1343-1353 [DOI 10.1002/btpr.1792].
- [6] Calliari C.M. and Go' mez R.J.H.C., "Extraction of pectin from orange peel waste (*Citrus sinensis*) by citric acid", In: *Proceedings of the XIX Brazilian Congress on Food Science and Technology*, 2004.
- [7] Virk B.S. and Sogi D.S., "Extraction and characterization of pectin from apple pomace (*Malus pumila* Cv amri) peel waste", *International Journal of Food Properties*. Vol 7, 2004, pp. 1–11.
- [8] Bensch M., Selbach B., Hubbuch H., "High throughput screening techniques in downstream processing: Preparation, characterization and optimization of aqueous two-phase systems", *Chemical Engineering Science*. Vol 67(7), 2007, pp. 2011-2021.
- [9] Nguyen Van Mui, *Practical Biochemistry*, Ha Noi National University Publishing House, 2001, pp.60-61 (Vietnamese).
- [10] Food Chemicals Codex, National Academy of Sciences Washington, 1972, pp. 580-581.

# Unknown Input Based Observer Design for Wind Energy Conversion System with Time-Delay

Van-Phong Vu

*Department of Automatic Control,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
phongvv@hcmute.edu.vn*

Wen-June Wang

*Department of Electrical Engineering,  
National Central University  
Zhongli, Taiwan  
wjwang@hcmute.edu.vn*

Van-Thuyen Ngo

*Department of Automatic Control,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thuyen.ngo@hcmute.edu.vn*

Dinh-Nhon Truong

*Department of Automatic Control,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
nhontd@hcmute.edu.vn*

Pei-Jun Lee

*Department of Electrical Engineering,  
National Chi-Nan University,  
Nantou, Taiwan  
pjlee@ncnu.edu.tw*

Ton Duc Do

*Department of Robotics and  
Mechatronics  
Nazarbayev University  
Astana Z05H0P9, Kazakhstan  
doduc.ton@nu.edu.kz*

Van-Ngoc Tong

*Faculty of Electrical and Electronics Engineering,  
Vietnam-Singapore Vocational College  
Binh Duong, Vietnam  
tongvanngocspkt@gmail.com*

**Abstract**—A new approach to design a nonlinear observer for the Wind Energy Conversion System (WECSs) with the existence of a delay time state variable is proposed in this paper. The WECSs are modeled under a polynomial framework that allows us to apply the powerful methodology of the linear system to synthesize the observer. Unlike the previous paper, the delay time in this paper is arbitrary and unnecessary to satisfy the bound constraints or the time delay must be constant. Additionally, during operating the WECSs, the wind is taken into account the unknown disturbance that affects to the performance of WECSs. In this paper, the unknown input method is employed to design the observer which can eliminate the influences of the delay time parts and disturbance; as well as estimate the unknown states asymptotically. Based on the Lyapunov theory and Sum-Of-Square (SOS) technique, the conditions to synthesize the nonlinear observer are obtained in the main theorems. The simulation results are also provided to show the merit of the proposed method.

**Keywords**—WECSs, Polynomial System, Unknown Input Method, Observer Design, SOS.

## I. INTRODUCTION

Recently, pollution increasingly becomes a serious issue in whole the world. Due to this reason, finding the solution to solve the polluted problem is a pressing issue that is received a great deal of attention from researchers. The major part of pollution may come from using fossil fuels such as oil and coal. To deal with this issue, recently, renewable energy such as wind energy and solar energy is one of the best options. There are plenty of studies focused on WECS in the past decades [1]-[8]. For example, in paper [1], an adaptive controller based on  $L_1$  technique was developed to track the maximum power point for the Wind Energy Conversion

Systems in which the parameter uncertain and the unknown disturbance influenced to the system. The sliding mode controller was designed for the low-power wind energy conversion system to obtain the maximum power generation of WECSs as well as maximize the power injection of the grid-connected power converter [3]. In paper [6], an active disturbance rejection controller was investigated for the WECSs to eliminate the effects of parametric uncertainties and the disturbance.

In practice, there exist many systems in which some state variables of these systems are unable to measure by sensor or cost to buy sensors is very high, therefore, observer synthesis is an alternative method to obtain the information of these states. Observer design increasingly plays an important role in the control field. There are a lot of studies concentrating on observer design published in recent years. For instance, the methods for designing an observer for the T-S fuzzy system have been studied in papers [9]-[11]. Regarding observer design for WECSs, several papers were published in recent years such as [12] and [13] where the observer is designed to estimate both unknown states and the disturbance of the WECSs.

Additionally, in reality, many systems are impacted by the time delay that will make the controller and observer synthesis much more difficult. In order to overcome this difficulty, a lot of previous papers focusing on controller and observer synthesis for the system with the existence of the time delay have been studied in [14]-[18]. For example, a method to design a controller to stabilize the T-S fuzzy time-delay system was introduced [14], in which the conditions for controller design was relaxed by adding a slack variable. In paper [15], authors have been proposed a method to design an observer-based controller relied on H-infinite to

eliminate the effects of the delay time and disturbance. Regarding observer/controller design for the WECSs with effects of time delay, according to our literature review, a few works are taking into account this problem [18]-[19]. The controller and observer were synthesized for WECSs with the existence of time delay terms in [18] and [19], respectively. However, the delay time in the previous papers [14]-[19] must be constant [15],[16],[18] or satisfy the bounded constraints [14], [17], [19]. Otherwise, the methods in these papers are failed to design controller/observer for the time-delay WECSs. Besides, in WECS, the wind is considered as a disturbance that will degrade the performance of the system. To overcome these drawbacks, the objective of this paper is to propose a method based on the unknown input method for designing an observer to estimate unknown states of the time-delay WECSs and eliminate the impacts of the time delay parts and disturbance as well. It should be noted that the time delay in this work is arbitrary and unnecessary to fulfil the bounded constraints. The WECSs, in this paper, is modeled in term of polynomial system [19]-[24].

*Notations:*  $\Pi > 0$  ( $\Pi < 0$ ) stands for a positive (negative) definite matrix.  $\Pi^T$  and  $\Pi^{-1}$  indicate the transpose and inverse matrix.  $I$  is defined for the identity matrix.

## II. MATHEMATICAL MODEL OF WECSs WITH TIME DELAY

Wind energy is transformed into power by the following formula.

$$P_a = \frac{1}{2} \rho \pi R^2 C_p(\lambda, \beta) v^3 \quad (1)$$

where  $\rho$  is the air density,  $v$  is the wind speed,  $R$  is the rotor radius of WT, and  $C_p(\lambda, \beta)$  defined as the power coefficient of WECSs is a nonlinear function which is dependent on the pitch angle  $\beta$  of the blades and the tip-speed ratio  $\lambda$ . The tip-speed ratio  $\lambda$  is expressed as follows

$$\lambda = \frac{\omega_t R}{v} \quad (2)$$

in which  $\omega_t$  is indicated the rotor speed. The aerodynamic torque  $T_a$  is computed

$$T_a = \frac{P_a}{\omega_t} = \frac{1}{2} \rho \pi R^3 C_q(\lambda, \beta) v^2 \quad (3)$$

where  $C_q(\lambda, \beta) = C_p(\lambda, \beta)/\lambda$  is the torque coefficient.

The ratio between speed and torque of the turbine side and generator side is determined as follows,

$$n_{gb} = \frac{\omega}{\omega_t} = \frac{T_a}{T_{gs}} \quad (4)$$

where  $n_{gb}$ ,  $T_{gs}$  and  $\omega$  are the gearbox ratio, equivalent aerodynamic torque of the generator, and mechanical angular speed of the generator. The PMSG's mathematical model is presented in the following equation [13]

$$\begin{cases} \frac{d\omega}{dt} = -\frac{B_v}{J} \omega - \frac{1}{J} T_e + \frac{1}{J n_{gb}} T_a \\ \frac{dT_e}{dt} = -\frac{R_s}{L} T_e - PK \omega i_d - \frac{\psi_m PK}{L} \omega + \frac{K}{L} v_q \\ \frac{di_d}{dt} = -\frac{R_s}{L} i_d + \frac{P}{K} \omega T_e + \frac{1}{L} v_d \end{cases} \quad (5)$$

From (5), it is rewritten

$$\begin{bmatrix} \dot{\omega} \\ \dot{T}_e \\ \dot{i}_d \end{bmatrix} = \begin{bmatrix} -\frac{B_v}{J} & -\frac{1}{J} & 0 \\ -\frac{\psi_m PK}{L} & -\frac{R_s}{L} & PK\omega \\ 0 & \frac{P}{K}\omega & -\frac{R_s}{L} \end{bmatrix} \begin{bmatrix} \omega \\ T_e \\ i_d \end{bmatrix} + \begin{bmatrix} \frac{0}{L} & \frac{0}{L} \\ \frac{K}{L} & 0 \\ 0 & \frac{1}{L} \end{bmatrix} \begin{bmatrix} v_q \\ v_d \end{bmatrix} + \begin{bmatrix} \frac{1}{J n_{gb}} \\ 0 \\ 0 \end{bmatrix} T_a \quad (6)$$

where  $K = 3/2 \psi_m P$ ,  $i_d$  is the stator currents in  $d$ -axis;  $v_d$  and  $v_q$ , are the stator voltages in  $d$ -axis and  $q$ -axis, respectively, which are taken into consideration as the inputs of the system, and  $T_e$  is the electromagnetic torque. Table 1 describes the definition and value of the parameters of the system WECS (6).

Then the system (6) is modeled under the framework of the polynomial system as follows

$$\begin{cases} \dot{x} = A(\omega)x + Bu + DT_a \\ y = Cx \end{cases} \quad (7)$$

where

$$x = [\omega \quad T_e \quad i_d]^T, u = [v_q \quad v_d]^T, y_1 = \omega$$

$$A(\omega) = \begin{bmatrix} -\frac{B_v}{J} & -\frac{1}{J} & 0 \\ -\frac{\psi_m PK}{L} & -\frac{R_s}{L} & PK\omega \\ 0 & \frac{P}{K}\omega & -\frac{R_s}{L} \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ \frac{K}{L} & 0 \\ 0 & \frac{1}{L} \end{bmatrix}.$$

$$C = [1 \quad 0 \quad 0] \text{ and } D = \begin{bmatrix} \frac{1}{J n_{gb}} & 0 & 0 \end{bmatrix}^T$$

From eq. (7), the aerodynamic torque  $T_a$  is taken into consideration as an unknown disturbance.

It is assumed that the WECSs system (7) is impacted by the time-delay terms in the following formula

$$\begin{cases} \dot{x}(t) = A(\omega)x(t) + A_d(\omega)x(t - \tau(t)) + Bu + DT_a \\ y = Cx \end{cases} \quad (8)$$

where

$A_d(\omega)$  is the matrix of the delay time state variables;  $x(t - \tau(t))$  is the vector of delay time state variables, and  $\tau(t)$  is the arbitrary time-varying delay.

In this paper, we assume that only mechanical angular speed of the generator  $\omega$  is measured by sensors. The electromagnetic torque  $T_e$  and current  $i_d$  are unknown. However, to design the controller to control the WECS and monitor states of WECS, these variable states are necessary to measure. Due to this reason, it inspires us to proposed a new approach to design the observer for estimating both the unknown states as well as eliminating the impacts of the time delay and disturbance Ta

**Assumption 1:** Assume that the matrix of time delay  $A_d(\omega)$  matrix satisfies the matching condition  $A_d(\omega) = D\Theta(\omega)$  in which  $\Theta(\omega)$  is an arbitrary polynomial matrix with appropriate dimensions.

**Remark 1:** The system (8) is impacted by the state delay time and the electromagnetic torque  $T_a$  that is considered as the disturbance. The existence of delay time and disturbance  $T_a$  will make the observer synthesis for the system (8) much more complicated. The time-delay in this system is an arbitrary

**Remark 2:** It should be noted that unlike the previous papers [14]-[18], the time delay  $\tau(t)$  in this paper is arbitrary and unnecessary to satisfy any constraint such as bounded constraint or must be constant. Thus, the proposed method in this paper is more relaxed.

### III. OBSERVER DESIGN

In this section, an observer will be synthesized to estimate the unknown states and eliminate the impacts of the disturbance as well as time-delay states.

Let us take into consideration the observer form for WECS model (8) as follows

$$\begin{cases} \dot{z}(t) = N(\omega)z(t) + Gu(t) + L(\omega)y(t) \\ \hat{x}(t) = z(t) - Ey(t) \end{cases} \quad (9a) \quad (9b)$$

in which  $\hat{x}(t) \in \mathbb{R}^{3 \times 1}$  are the estimation of  $x$  and  $z(t) \in \mathbb{R}^{3 \times 1}$  is the state variable vector of the observer (9).  $N(\omega) \in \mathbb{R}^{3 \times 3}$ ,  $G \in \mathbb{R}^{3 \times 2}$ ,  $L(\omega) \in \mathbb{R}^{3 \times 2}$ ,  $E \in \mathbb{R}^{3 \times 2}$  are the parameters of the observer (9) that will be determined later.

Because, the system matrices of the system (8) are polynomial matrices, therefore the conditions for designing observer is expressed in term of Sum-Of-Square (SOS). To solve the conditions for synthesizing the observer in the next steps, two following propositions are needed.

**Proposition 1 [20]:**  $\Omega(\varepsilon(t))$  is called a Sum-Of-Square (SOS) if  $\Omega(\varepsilon(t)) = \sum_{i=1}^n [v_i(\varepsilon(t))]^2$ , in which  $v_i(\varepsilon(t))$  is a polynomial in  $\varepsilon(t)$ . If  $\Omega(\varepsilon(t))$  is a SOS then we can conclude that  $f(x(t)) \geq 0$ , however, the converse is not guaranteed.

**Proposition 2 [20]:** Taken into consideration a square polynomial symmetric matrix  $\Theta(\varepsilon)$  with dimension  $n \times n$  and a vector  $v \in \mathbb{R}^n$  is not dependent on  $\varepsilon$ ,  $\Theta(\varepsilon) \geq 0$ , if only if  $v^T \Theta(\varepsilon) v$  is presented in term of an SOS.

**Theorem 1:** The estimation errors of the system (8) with the observer (9) converge to zero asymptotically if there exist matrices  $N(\omega)$ ,  $G$ ,  $L(\omega)$ ,  $F$  and a symmetric matrix  $P$  such that the following conditions hold:

$$KA(\omega) - N(\omega)K - L(\omega)C = 0 \quad (10)$$

$$KB - G = 0 \quad (12)$$

$$KD = 0 \quad (13)$$

$$v_1^T (Q - \varepsilon_1 I) v_1 \text{ is SOS} \quad (14)$$

$$-v_2^T (N^T(\omega)Q + QN(\omega) - \varepsilon_2(\omega)I) v_2 \text{ is SOS} \quad (15)$$

where

$$K = I + EC$$

$v_1, v_2$  is the vector independent on  $\omega$ ,  $\varepsilon_1$  is positive scalar and  $\varepsilon_2(\omega) > 0$  with  $\omega \neq 0$ .

**Proof:**

Based on Assumption 1, the system (8) is rewritten as follows

$$\begin{cases} \dot{x}(t) = A(\omega)x(t) + Bu + D(\Theta(\omega)x(t - \tau(t)) + T_a) \\ y = Cx \end{cases} \quad (16)$$

Let us define:

$$\phi(\omega) = (\Theta(\omega)x(t - \tau(t)) + T_a)$$

Then the system (16) becomes

$$\begin{cases} \dot{x}(t) = A(\omega)x(t) + Bu + D\phi(\omega) \\ y = Cx \end{cases} \quad (17)$$

The estimation error is formulated as follows

$$\begin{aligned} e(t) &= x(t) - \hat{x}(t) = x(t) - z(t) + Ey(t) \\ &= (I + EC)x(t) - z(t) = Kx(t) - z(t) \end{aligned} \quad (18)$$

where

$$K = I + EC \quad (19)$$

The derivative of the estimation errors is computed by

$$\dot{e}(t) = K\dot{x}(t) - \dot{z} \quad (20)$$

From (8), (9a), (18) and (20), one can obtain

$$\begin{aligned} \dot{e}(t) &= K[A(\omega)x(t) + Bu + D\phi(\omega)] \\ &\quad - [N(\omega)z(t) + Gu(t) + L(\omega)y(t)] \\ &= N(\omega)e(t) + [KA(\omega) - N(\omega)K - L(\omega)C]x + \\ &\quad + (KB - G)u(t) + KD\phi(\omega) \end{aligned} \quad (21)$$

It is obvious that if the conditions (10)-(13) of Theorem 1 hold, then (21) becomes

$$\dot{e}(t) = N(\omega)e(t) \quad (22)$$

Select the Lyapunov function as follows

$$V(e(t)) = e^T(t)Qe(t) \quad (23)$$

If the condition (14) of Theorem 1 is fulfilled, it guarantees that the matrix  $Q$  is a positive definite matrix. It leads to  $V(e(t)) > 0$ .

From (23), it yields

$$\dot{V}(e(t)) = \dot{e}^T(t)Qe(t) + e^T(t)Q\dot{e}(t) \quad (24)$$

Combining (22) and (24) yields

$$\dot{V}(e(t)) = e^T(t)[N^T(\omega)Q + QN(\omega)]e(t) \quad (25)$$

From the equation (25), it is obvious that if the condition (15) of Theorem 1 holds, then  $\dot{V}(e(t)) < 0$ . It means that estimation error  $e(t)$  approach to zero asymptotically. The proof is completed.

It should be noted that (15) implies that

$$N^T(\omega)Q + QN(\omega) < 0 \quad (26)$$

However, the condition (26) is not a Polynomial Linear Matrix Inequality (PLMI), thus it is unable to solve this equation in the SOS Tool of MATLAB. To overcome this difficulty, the PBMI (polynomial Bilinear Matrix Inequality) is converted to the Polynomial Linear Matrix Inequality (PLMI).

**Theorem 2:** The estimated states approach the real states of the system (8) with the observer (9) if there exist the matrices  $N(\omega)$ ,  $G$ ,  $L(\omega)$ ,  $F$ , and symmetric matrix  $Q$  such that the following conditions are fulfilled

$$v_1^T(Q - \varepsilon_1 I)v_1 \text{ is SOS} \quad (27)$$

$$-v_2^T(\Xi(\omega) + \varepsilon_2(\omega)I)v_2 \text{ is SOS} \quad (28)$$

in which

$$C^T \bar{Z}^T(\omega) - [(I - D(CD)^{-1}C)A(\omega)]^T Q + \bar{Z}(\omega)C - Q[(I - D(CD)^{-1}C)A(\omega)] < 0 \quad (29)$$

$$\bar{Z}(\omega) = QZ(\omega) \quad (30)$$

$$K = I + EC$$

$v_1$  and  $v_2$  are the vectors that are independent on  $\omega$ .  $\varepsilon_1$  is the positive scalar and  $\varepsilon_2(\omega) > 0$  with  $\omega \neq 0$ .

The parameters of the observer (9) are computed as follows

$$E = -D(CD)^{-1} \quad (31)$$

$$N(\omega) = Z(\omega)C - KA(\omega) \quad (32)$$

$$L(\omega) = Z(\omega)(I - CE) + KA(\omega)E \quad (33)$$

$$G = KB \quad (34)$$

**Proof:** Because of page limitation, the proof of Theorem 2 is omitted.

#### IV. SIMULATION RESULTS

In this part, the polynomial observer is synthesized and the simulation with parameter in Table 1 is executed to express the successes of the proposed method.

In this simulation, on the basis of Assumption 1, the matrix of the time delay state  $A_d(\omega)$  is decomposed into

$$\text{matrices } D = \begin{bmatrix} 1 \\ Jn_{gb} \\ 0 \\ 0 \end{bmatrix} \text{ and } \Theta(\omega) = [\omega \quad \omega + 1 \quad \omega^2]. \text{ The}$$

time delay is chosen  $\tau(t) = 5(s)$ . It should be noted that this study focuses on the observer design for WECSs and controller design is out of the scope of this paper, therefore, in the simulation, the control signal is arbitrarily selected  $u = \sin(t)$ .

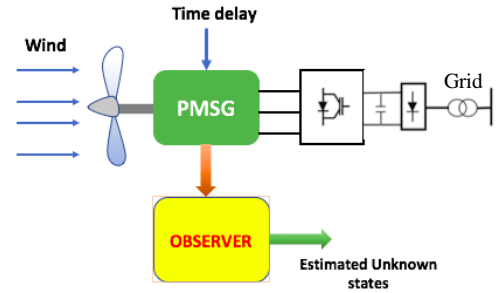


Fig.1. Structure of the system

TABLE 1. WECS PARAMETERS

Symbols	Parameters	Values	Units
$P_{rated}$	Rated power	5	kW
$P$	Pole pairs	14	-
$R_s$	Stator resistance	0.3676	$\Omega$
$L$	Stator inductance	3.55	mH
$\Psi_m$	Magnet flux linkage	0.2867	V.s/rad
$J$	Mechanical inertia	7.856	Kg.m <sup>2</sup>
$B_v$	Viscous friction coefficient	0.002	Kg.m <sup>2</sup> /s
$R$	Rotor radius	1.84	m
$\rho$	Air density	1.25	Kg.m <sup>3</sup>
$n_{gb}$	Gearbox ratio	1	-

Solving the conditions in Theorem 2 by SOS tools of Matlab obtains

$$N(\omega) = \begin{bmatrix} N_{11}(\omega) & 0 & 0 \\ N_{21}(\omega) & -103.549 & 84.2898\omega \\ N_{31}(\omega) & 2.325\omega & -103.6 \end{bmatrix}$$

where

$$N_{11}(\omega) = -0.597\omega^2 + 0.16 * 10^{-7}\omega + 0.686$$

$$N_{21}(\omega) = 0.145\omega^2 + 0.17 * 10^{-5}\omega - 62.41$$

$$N_{23}(\omega) = -0.59 * 10^{-7}\omega^2 + 0.298\omega + 0.269 * 10^{-7}$$



$$L(\omega) = \begin{bmatrix} 0 \\ -6.8073e + 03 \\ 0 \end{bmatrix}, F = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$$

$$G = 10^3 * \begin{bmatrix} 0 & 0 \\ 1.7 & 0 \\ 0 & 0.28 \end{bmatrix}.$$

Carrying out the simulation in Matlab, the obtained results are illustrated in Fig. 2.

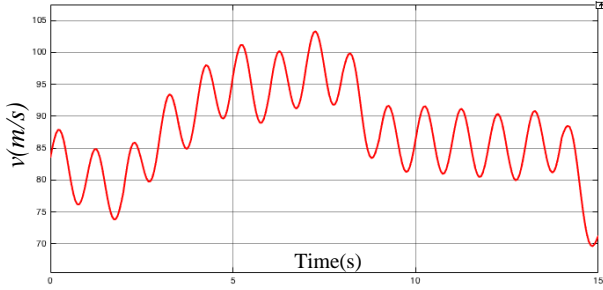


Fig.2. Wind wave form

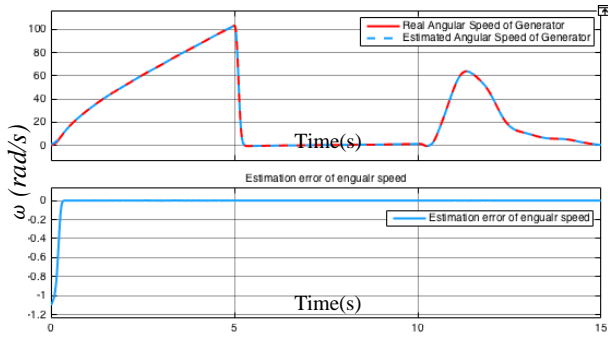


Fig. 3. Real angular speed of generator  $\omega$ , estimation  $\hat{\omega}$ , and estimation error  $\omega - \hat{\omega}$ ,

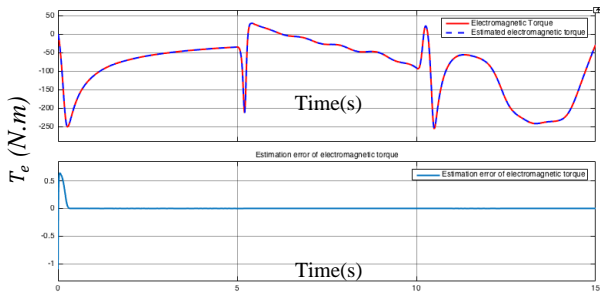


Fig. 4. Real Electromagnetic torque  $T_e$ , estimation  $\hat{T}_e$ , and estimation error  $T_e - \hat{T}_e$ ,

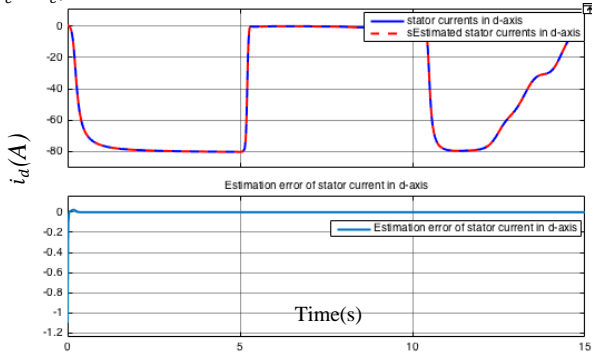


Fig. 5. Real stator current  $i_d$ , estimation  $\hat{i}_d$  and estimation error  $i_d - \hat{i}_d$ .

The simulation results from Figs. 3-5 express the real angular speed of generator  $\omega$ , electromagnetic torque  $T_e$ , and

stator current  $i_d$  are estimated asymptotically. The estimation errors  $e_\omega = \omega - \hat{\omega}$ ,  $e_{T_e} = T_e - \hat{T}_e$ , and  $e_{i_d} = i_d - \hat{i}_d$  approach zero asymptotically. The simulation results illustrated in Figs. 3-5 have shown that the polynomial observer of WECSs is successfully designed even with the existence of the time delay terms.

## V. CONCLUSION

A new approach to synthesize the observer to estimate the unmeasurable states has been investigated in this paper.

The WECS is impacted by the time delay parts in which the time delay is arbitrary and does not need to satisfy any constraint. Even with the influences of the time delay and disturbance, the proposed method still successfully estimates the unknown states. The conditions to build the observer are formulated in terms of the SOS framework in Theorem 1 and 2. The simulation results are provided to show the success of the proposed method.

## ACKNOWLEDGMENT

We would like to thank Ho Chi Minh City University of Technology and Education; and the Ministry of Sciences and Technology who have supported this paper under funding of project with grant number CT2019.04

## REFERENCES

- [1] H. Zhao, Q. Wu, C. N. Rasmussen, and M. Blanke, "L<sub>1</sub> Adaptive speed control of a small wind energy conversion system for maximum power point tracking," *IEEE Trans. Energy Conversion*, vol. 29, no. 3, pp. 576-584, 2014.
- [2] Z. Q. Jin, F. X. Li, X. Ma, and S. M. Djouadi, "Semi-definite programming for power output control in a wind energy conversion system," *IEEE Trans. Sustainable Energy*, vol. 5, no. 2, pp. 466-474, 2014.
- [3] Y. Yang, K. T. Mok, S. C. Tan, and S. Y. R. Hui, "Nonlinear Dynamic Power Tracking of Low-Power Wind Energy Conversion System," *IEEE Trans. Power Electronics*, vol. 30, no. 9, pp. 5223 - 5236, 2015.
- [4] M. Moness and A. M. Moustafa, "A Survey of Cyber-Physical Advances and Challenges of Wind Energy Conversion Systems: Prospects for Internet of Energy," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 134 - 145, 2016.
- [5] Y. J. Ma, L. Tao, X. S. Zhou, and X. Q. Shi, "Analysis and Control of Fault Ride-Through Capability Improvement for Wind Energy Conversion System Using Linear Active Disturbance Rejection Control With Correction Link," *IEEE Access*, vol. 8, pp. 73816 - 73827, 2020.
- [6] S. Das and B. Subudhi, "A Two Degree of Freedom Internal Model based Active Disturbance Rejection Controller for a Wind," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2019, DOI: 10.1109/JESTPE.2019.2905880. To be published.
- [7] J. Hussain, and M. K. Mishra, 'An Efficient Wind Speed Computation Method Using Sliding Mode Observers in Wind Energy Conversion System Control Applications,' *IEEE Transactions on Industry Applications*, vol 56, no. 1, pp. 730 - 739, 2020.
- [8] A. Sattar, A. Al-Durra, C. Caruana, M. Debouza, and S. M. Mueen, 'Testing the Performance of Battery Energy Storage in a Wind Energy Conversion System,' *IEEE Trans. Industry Applications*, vol. 56, no. 3, pp. 3195-3206, 2020.
- [9] A. Chibani, M. Chadli, M.M. Belhaouane, et al., "State estimation of unknown input polynomial systems: a sum of squares approach," *The 22nd Mediterranean Conf. on Control and Automation (MED)*, Palermo, Italy, 2014, pp. 1225-1230.
- [10] V. P. Vu and W. J. Wang, "Observer design for a discrete time T-S fuzzy system with uncertainties. The 2015 IEEE International Conference on Automation Science and Engineering (CASE), Gothenburg, Sweden, 2015, pp. 1262-1267.

- [11] V. P. Vu and W. J. Wang, "State/Disturbance observer synthesis for T-S fuzzy system with the enlarge class of disturbances," *IEEE Trans. Fuzzy Syst.* vol. 26, no. 6, pp. 3645-3659, 2018.
- [12] A. V. Le, T. D. Do, "High-order observers-based LQ control scheme for wind speed and uncertainties estimation in WECSs," *Optimal Control and Application Methods*, vol. 39, no. 5, pp. 1818-1832, 2018.
- [13] V. P. Vu and T. D. Do, "A Novel Nonlinear Observer Based LQ Control System Design for Wind Energy Conversion Systems with Single Measurement", *Wind Energy*, 2019.
- [14] S. H. Tsai, Y. A. Chen, and J. C. Lo, "A Novel Stabilization Condition for A Class of T-S Fuzzy Time-Delay Systems," *Neurocomputing*, vol. 175, pp. 223-232, 2016.
- [15] C. Lin, Q. G. Wang, T. H. Lee, and Y. He, "Design of Observer-Based  $\infty$ Control for Fuzzy Time-Delay Systems," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 2, pp. 534-543, 2008.
- [16] M. Han, H.K. Lam, Y. D. Li, F. C. Liu, C. Z. Zhang, "Observer-based control of positive polynomial fuzzy systems with unknown time delay," *Neurocomputing*, vol. 349, pp. 77-90, 2019.
- [17] H. R. Karimi and Mohammed Chadli, "Design of Robust Observer for T-S Fuzzy Time-Delayed Systems Subject to Unknown Inputs," *Proceedings of 2013 International Conference on Fuzzy Theory and Its Application*, Taipei, Taiwan, Dec. 6-8, 2013, pp. 100-104.
- [18] X. Wang and M. J. Alden, "Resilient and Robust Control of Time-Delay Wind Energy Conversion System," *ASME J. Risk Uncertainty Part B*, vol. 3, no.1, pp. 011005.
- [19] A. Chibani, M. Chadli, M.M. Belhaouane, et al., "State estimation of unknown input polynomial systems: a sum of squares approach," *The 22nd Mediterranean Conf. on Control and Automation (MED)*, Palermo, Italy, 2014, pp. 1225-1230.
- [20] M.S.B.M. Saat, "Controller synthesis for polynomial discrete-time systems," *Ph.D. dissertation*, the University of Auckland, 2013
- [21] A. Papachristodoulou, J. Anderson, G. Valmorbida, et al., "SOS tools sum of squares optimization toolbox for MATLAB user's guide version 3.00," 2013.
- [22] S. Saat, S.K. Nguang, A.M. Darsono, et al. "Nonlinear  $H_\infty$  feedback control with integrator for polynomial discrete-time systems," *J. Franklin Inst.*, vol. 351, pp. 4023-4038, 2014.
- [23] V. P. Vu and W. J. Wang, "Observer-based controller synthesis for uncertain polynomial systems", *IET Control Theory and Applications*, vol.12, no.1, pp. 29-37, 2018.
- [24] V. P. Vu, W. J. Wang, H. C. Chen, and J. M. Zurada, "Unknown Input Based Observer Synthesis for a Polynomial T-S Fuzzy Model System with Uncertainties," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1447 - 1458, 2018.

# Analysis of Flexible Pavements Comprised of Conventional and High Modulus Asphalt Concrete Subjected to Moving Loading using Linear Viscoelastic Theory

H.T. Tai Nguyen

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tainht@hcmute.edu.vn

Thanh-Nhan Phan

Department of Civil Engineering  
Faculty of Architecture, Thu Dau Mot  
University  
Binh Duong Province, Vietnam  
nhanpt@tdmu.edu.vn

Tien-Tho Do

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thodt@hcmute.edu.vn

Duy-Liem Nguyen

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
liemnd@hcmute.edu.vn

Vu-Tu Tran

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
tutv@hcmute.edu.vn

**Abstract**— In order to prevent distresses in asphalt pavement, the analysis of pavement structure should be as close as possible to its real behavior. However, in actual standard for designing asphalt pavement, viscoelastic behavior of asphalt mixtures and effects of moving vehicles are still not considered. In this paper, the authors aim at applying an analytical approach for analyzing the linear viscoelastic behavior of asphalt pavements comprised of conventional and high modulus asphalt concrete caused by a moving loading. It is shown that the effects of moving load give rise to an increase in tensile stress at the pavement surface and at the bottom of asphalt layers, which can be a reason for premature cracks in flexible pavement. The results also show that the use of high modulus asphalt concrete layers gives rise to a significant reduction of tensile stress and strain at the pavement surface as well as at the bottom of asphalt layer.

**Keywords**—Linear viscoelastic, moving load, complex modulus, creep compliance, multi-layered half-space

## I. INTRODUCTION

In most countries, asphalt pavement represents more than 90% of road transport system. As a result, a longer in-service life of flexible pavements will contribute to a reduction in transportation time and cost for the national economy. Although methods for design and construction of long-lasting pavement, or perpetual pavement, has been developed since the 1960s, they are still among most attracting subjects of study for many local authorities, researchers, and practitioners around the world [1,2]. A perpetual pavement structure is dimensioned thick enough and the asphalt concrete materials is well chosen so that rutting distress and bottom-up cracking can be eliminated while top-down cracking and other types of distress such as wearing (if occurring) are localized only in the top layer. With this design concept, only the top layer is necessary to maintenance and repair activities while lower layers have very long in-service life and do not need any maintenance during the lifetime of the pavement.

In Vietnam, the pavement of national highways comprises normally two asphalt layers of 12-15 cm in thickness [3,4]. Because of very thin asphalt pavement, the 60/70 pen bitumen, which is a soft bitumen, is widely used in the country to produce asphalt mixture (hereafter denoted as conventional asphalt concrete, or simply AC) to prevent fatigue cracking. With hot climatic conditions and overloaded vehicles, rutting distress has been a severe problem in the country [5]. As a result of many attempts made in the country, rutting distress can be effectively predicted and prevented thanks to prediction methods, improvement in material performance and construction techniques, e.g. Refs. [5,6,7,8]. However, the community is facing another problem once the pavement is stiff enough to prevent rutting. That is fatigue cracking including bottom-up and top-down cracking occurring more and more frequently than before. The following figure (Fig. 1) shows fatigue cracking pattern of a pavement comprised of an SBS-modified and a conventional asphalt layer just after two years in service. Therefore, advanced methods for pavement analysis and pavement distress prediction should be considered and applied to prevent such type of distress for upcoming flexible pavements.



Fig. 1. Fatigue cracks and pothole in National Highway 1A, located near Thu Duc District, Ho Chi Minh City, Vietnam [9].

Asphalt concrete shows both viscous and elastic behavior, which depends strongly on loading rate (frequency or time) and temperature. Linear viscoelastic (LVE) theory has been successfully used for longtime ago to characterize the behavior of asphalt concrete materials [10,11,12,13]. Therefore, the analysis of flexible pavements using LVE theory certainly gives better results than does the elastic theory. Unfortunately, the actual national standard design code [14], which is based on elastic theory, does not consider the effect of loading rate in pavement analysis and the pavement distress prediction method is no more adaptable to actual traffic conditions. Owing to the lack of testing devices, little study on LVE characterization of asphalt mixtures and LVE analysis of pavement structures has been performed in the country, e.g. Refs. [15,16,17,18]. In addition, the effect of moving axle load on the observation response of the pavement is still not considered by local researchers.

This study aims at investigating the effect of moving loading on the response of a flexible pavement comprised of conventional and high modulus asphalt concrete (HMAC). We will focus on the deflection defined as the maximal displacement of the pavement surface, stress and strain at the surface and at the bottom of second asphalt concrete layer. To achieve this goal, the moving loading is decomposed into a sequence of stationary loading as proposed by Levenberg [19] and Boltzmann's superposition principle was used to calculate the pavement responses at the observation point due to a moving loading [19,20]. The results of this study will contribute to the improvement of analysis and design of flexible pavement comprised of conventional AC and HMAC.

## II. THEORETICAL FORMULATION

### A. Characterization of viscoelastic behavior of asphalt mixtures

The viscoelastic properties of asphalt mixtures can be mathematically presented by a function of complex modulus or complex compliance which is defined in the frequency domain. The stress strain relationship at a material point can be written as:

$$\sigma^* = E^*(i\omega) \cdot \varepsilon^* \quad (1)$$

where  $\varepsilon^*$  is complex number representing the sinusoidal input of strain,  $\sigma^*$  is complex number representing the sinusoidal response of stress and  $E^*(i\omega)$  is the corresponding complex modulus,  $\omega$  is the frequency of the applied input and  $i$  is the imaginary number. The complex compliance is defined as the inverse of complex modulus:

$$D^*(i\omega) = 1 / E^*(i\omega). \quad (2)$$

In the time domain, the behavior of a viscoelastic material can be presented by a function of relaxation modulus or creep compliance, in which the relaxation modulus ( $E(t)$ ) is defined as the stress response at a material point due to a step constant strain of unity  $\varepsilon(t) = 1$ , for  $t > 0$ :

$$\sigma(t) = E(t) \cdot 1 \quad (3)$$

and the creep compliance ( $D(t)$ ) is defined as the strain response due to a step constant stress of unity  $\sigma(t) = 1$ , for  $t > 0$ :

$$\varepsilon(t) = D(t) \cdot 1 \quad (4)$$

In order to characterize the full behavior of the material, the complex modulus/compliance or the relaxation modulus/creep compliance must be determined in a very wide range of frequency/loading time, varying from very small to very high values, e.g.  $10^{-8}$  to  $10^8$ . Due to machine limits of testing devices, one cannot perform tests either at very small values or very high values of frequency/loading time. However, we can construct the curve of the viscoelastic properties in a wide range of frequency/loading, called the master curve thanks to the time-temperature correspondence principle. In effect, the viscoelastic properties of the material at a frequency  $\omega$  or loading time  $t$  and temperature  $T$  is equivalent to those at frequency  $\omega'$  or loading time  $t'$  and temperature  $T'$ :

$$E^*(\omega', T') = E^*(\omega, T); E(t', T') = E(t, T) \quad (5)$$

Let choose a value of temperature as the reference temperature — say  $T_{ref}$ , the viscoelastic properties at small values of frequency or high loading time can be determined by performing tests at temperatures higher than  $T_{ref}$ . Similarly, the viscoelastic properties at high values of frequency or small loading time can be determined by performing tests at temperatures lower than  $T_{ref}$ . As a result, one must perform several tests at different temperatures and at each temperature, the frequency/loading time is swept in a wide range depending on the limit of test devices so as to construct a one single curve representing the viscoelastic properties in a full range of frequency/loading time of interest. The ratio of  $\omega'$  with respect to  $\omega$  at a temperature  $T$ , called the shift factor  $a_T(T)$ , is also an intrinsic property of viscoelastic material.

For asphalt mixtures, the characterization of viscoelastic properties in frequency domain is easier to perform than in time domain. The dynamic modulus  $|E^*|$  and phase angle  $\varphi$  are widely used and is defined as:

$$\begin{aligned} |E^*| &= \sqrt{E'^2 + E''^2} \\ \cos(\varphi) &= E' / |E^*|; \sin(\varphi) = E'' / |E^*| \end{aligned} \quad (6)$$

where  $E'$  and  $E''$  are the real and imaginary part of the complex number  $E^*$ , respectively. The determination of dynamic modulus and phase angle can be performed following the instructions in AASHTO T342 standard. After the master curves of dynamic modulus and phase angle have been constructed, they must be predicted using mathematical or rheological models so that the viscoelastic behavior of material can be implemented in a calculation program for analysis of pavement structure. The rheological model of Huet-Sayegh [21,22], or its generalized version, 2S2P1D model [23] are well-known for their excellent fitting of

viscoelastic behavior of asphalt mixtures while using very few parameters (see Fig. 2).

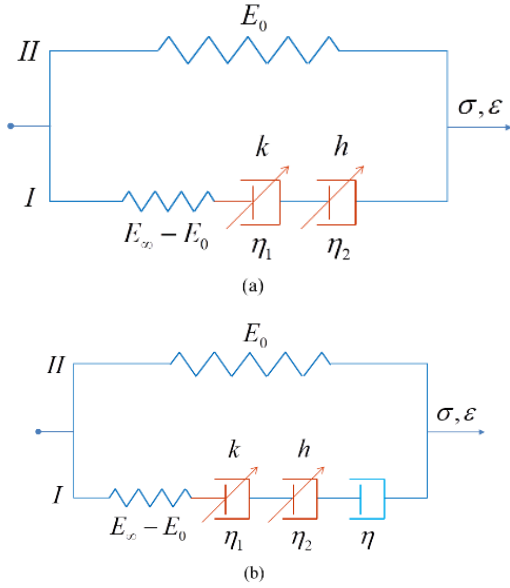


Fig. 2. Presentation of (a) Huet-Sayegh model and (b) 2S2P1D model.

The complex modulus of Huet-Sayegh model is:

$$E^*(\omega) = E_0 + \frac{E_\infty - E_0}{1 + \delta(i\omega\tau(T))^{-k} + (i\omega\tau(T))^{-h}} \quad (7)$$

and that of 2S2P1D model is (see Ref. [21]):

$$E^*(\omega) = E_0 + \frac{E_0 - E_\infty}{1 + \delta(i\omega\tau(T))^{-k} + (i\omega\tau(T))^{-h} + (i\omega\beta\tau(T))^{-1}} \quad (8)$$

where  $\omega$  (rad/s) is angular frequency,  $\delta = \frac{\tau(T) \cdot (E_\infty - E_0)}{\eta_1}$ ,

$0 < h < k < 1$  are model parameters,  $E_0, E_\infty$  are the value of  $E^*(\omega)$  when  $\omega\tau \rightarrow 0$  and  $\omega\tau \rightarrow \infty$  and  $\beta$  is a material parameter. In this paper, the LVE properties of conventional asphalt concrete and high modulus asphalt concrete which were characterized in Ref. [16] (and summarized in Tab. 1) are used to analyze the behavior of pavements subjected to moving loading.

TABLE I. HUET-SAYEGH PARAMETERS OF AC AND HMAC AT 20°C

Parameter	Mixture	
	AC	HMAC
$E_0$ (MPa)	150	425
$E_\infty$ (MPa)	37338	34227
$h$	0.32	0.21
$k$	0.8	0.59
$\delta$	5.72	2.07
$\tau$ (s)	0.10	0.63

## B. Interconversion between dynamic modulus and relaxation modulus

As previously mentioned, it is more convenient to determine the dynamic modulus and phase angle in the frequency domain. However, one also needs to know the behavior of the material in the time domain. In addition, most finite element programs need the relaxation modulus or creep compliance which is defined in the time domain as material properties input. Therefore, several interconversion methods for viscoelastic properties in the frequency and time domain have been proposed in the literature. In general, the relationship between these quantities is given in the form of Carson transformations:

$$\begin{aligned} \tilde{E}(s) &= s \int_0^\infty E(t) e^{-st} dt; \\ \tilde{D}(s) &= s \int_0^\infty D(t) e^{-st} dt \\ \tilde{E}(s) \tilde{D}(s) &= 1 \end{aligned} \quad (9)$$

$$\begin{aligned} E^*(\omega) &= \tilde{E}(s) \Big|_{s \rightarrow i\omega} \\ D^*(\omega) &= \tilde{D}(s) \Big|_{s \rightarrow i\omega} \end{aligned} \quad (10)$$

For practical use, the approximate analytical formula of Schapery and Park [24] is widely used for interconversion between viscoelastic properties in frequency and time domain and is shown in Eqs. (11)-(12):

$$E\left(t = \frac{1}{\omega}\right) = \frac{E'(\omega)}{\lambda'(\omega)}; \lambda'(\omega) = \Gamma(1-n) \cos\left(n \frac{\pi}{2}\right) \quad (11)$$

$$E(t) \cdot D(t) = \frac{\sin(n\pi)}{n\pi} \quad (12)$$

where  $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$  is Gamma function and  $n$  is the local slope of the curve of  $\log(E'(\omega))$  versus  $\log(\omega)$ . The derivation of relaxation modulus and creep compliance for HS and 2S2P1D models is presented in Ref. [18].

## C. Analytical approach for the problem of a viscoelastic multi-layered half-space subjected to a moving load

Let consider a multi-layered half-space comprised of asphalt concretes layers, unbound aggregate material layers (UGM) and a road base of infinite thickness as described in Fig. 2. The LVE properties of  $i^{\text{th}}$  asphalt concrete layer are represented by a creep compliance  $D_i(t)$  and Poisson ratio  $\nu_i$  while the elastic properties of  $i^{\text{th}}$  UGM layer and the road base are represented by an elastic modulus  $E_i$  and Poisson ratio  $\nu_i$ . The interaction between layers is considered as fully bonded, i.e. no relative slip or detachment is allowed. A Cartesian coordinate is defined as in the figure for locating the moving load, of which the direction of the moving load is from left to right along the  $x$ -axis.



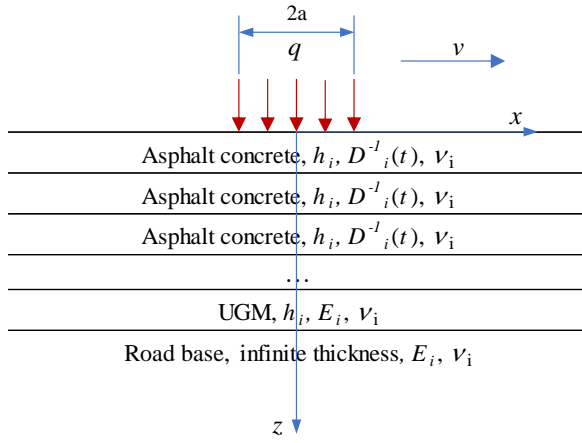


Fig. 3. Illustration of a general pavement structure

The applied load is a uniformly distributed force  $q$  moving along the  $x$ -axis, representing the application of one-half of an axle load as indicated in Fig. 3. The contact area is assumed to be a circle of diameter  $2a$ .

In case  $v=0$ , the load is stationary and let  $t_L$  is the loading time such that  $q(t) \neq 0$ , for  $t \leq t_L$  and  $q(t)=0$ , for  $t > t_L$ . The viscoelastic solution to the problem can be approximated using the Boltzmann's superposition principle [20]

$$R^{ve}(\mathbf{x}, t) = \int_0^t R_H^{ve}(\mathbf{x}, t-\tau) dI(\tau) \quad (13)$$

where  $R^{ve}(\mathbf{x}, t)$  is the LVE response at coordinates  $\mathbf{x}$  and time  $t$ ,  $R_H^{ve}(\mathbf{x}, t-\tau)$  is the LVE response of the pavement to a step input function at time  $t-\tau$ , and  $dI(\tau)$  is the variation of the input at time  $\tau$ . The discretized form of Eq. (13) can be written as:

$$R^{ve}(\mathbf{X}, t_i) = \sum_{j=0}^i R_H^{ve}(\mathbf{X}, t_i - t_j) \Delta I(t_j) \quad (14)$$

where  $R_H^{ve}(\mathbf{X}, t_i - t_j)$  is effectively the solution of a quasi-elastic problem using the relaxation modulus at time  $t_i - t_j$ . As a result, the viscoelastic solution can be approximated from a sequence of quasi-elastic solutions, which can be solved by using Burmister multi-layered elastic theory [25].

In case  $v \neq 0$ , the load is approaching the observation point ( $x = x_0$ ) from the left and overpassing this point. We cannot calculate the entire path of the moving load because of extremely high computation cost. Therefore, we will consider a path of the moving load of finite length  $L$ . The moving path starts from a point of coordinates ( $x = -L/2, y = 0$ ) and finish at point ( $x = +L/2, y = 0$ ). Levenberg suggests that  $L = 10m$  is enough to represent the real behavior of pavement due to a moving load of radius 150 mm [19]. In this paper,  $L = 100a$  is used. Let discretize the influencing path length  $L$  by  $N$  elements as in Fig. 4.

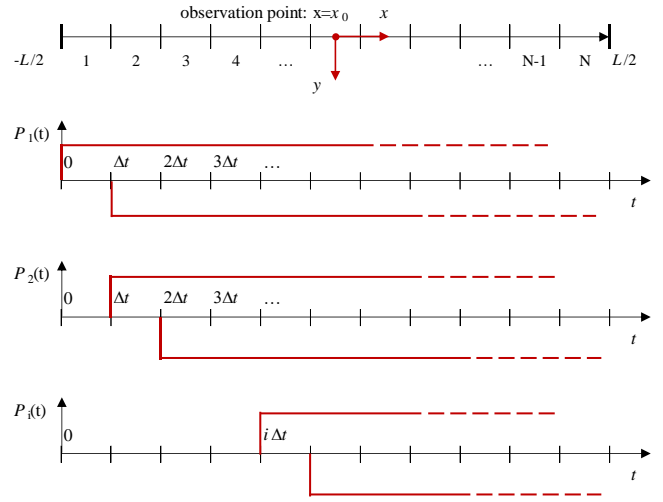


Fig. 4. Illustration of a pavement structure in general

The length of each element is therefore  $\Delta x = L/N$  and the corresponding loading time that the moving load applies to the pavement is  $\Delta t = L/(N \cdot v)$ . Let  $P_i(t)$  be the equivalent applying force history of the moving load when it is located at element  $i$ . It is more convenient to split  $P_i(t)$  into a loading and unloading step function as proposed by Levenberg [19] and illustrated in Fig 4. The delay time of the unloading part is equal to the loading time  $\Delta t$ . Once again, by making use of the Boltzmann's superposition principle, one can derive a formula for calculating the viscoelastic response of the pavement subjected to a moving load as follows.

Let defined  $x_i$  is the  $x$ -coordinate of the moving load when it is at  $i^{\text{th}}$  element,  $x_i = -[(N-3)/2 + i]\Delta x$ . The response at the observation point depends not only on the equivalent load at current position of the moving load but also on that at previous positions. When the moving load is at element 1, the viscoelastic response at the observation point is:

$$R^{ve}(1) = R_H^{ve}(D^{-1}(t_1), x_1) \quad (15)$$

where  $R_H^{ve}(D^{-1}(t_1), x_1)$  is the LVE response due to loading portion of  $P_1(t)$  during the loading time  $t_L = 1 \cdot \Delta t$ . After that, the loading moves from element 1 to element 2, the LVE response at the observation point is:

$$R^{ve}(2) = R_H^{ve}(D^{-1}(t_1), x_2) + R_H^{ve}(D^{-1}(t_2), x_1) - R_H^{ve}(D^{-1}(t_1), x_1) \quad (16)$$

where  $R_H^{ve}(D^{-1}(t_1), x_2)$  is the LVE response at  $(x_0, y_0)$  due to the loading portion of  $P_2(t)$  during the loading time  $t_L = 1 \cdot \Delta t$ ,  $R_H^{ve}(D^{-1}(t_2), x_1)$  is the LVE response  $(x_0, y_0)$  due to the loading portion of  $P_1(t)$  during the loading time  $t_L = 2 \cdot \Delta t$ , and  $-R_H^{ve}(D^{-1}(t_1), x_1)$  is the LVE response at  $(x_0, y_0)$  due to the unloading portion of  $P_1(t)$  during the loading time  $t_L = 1 \cdot \Delta t$ . For the sake of simplicity, Eq. (16) can be rewritten as:

$$R^{ve}(2) = R_H^{ve}(1, 2) + R_H^{ve}(2, 1) - R_H^{ve}(1, 1) \quad (17)$$

In a similar way, the LVE response at the observation point when the moving load is at position  $x_i$  is given by:

$$R^{ve}(i) = R_H^{ve}(1, i) + \sum_j^{i-1} [R_H^{ve}(i-j+1, j) - R_H^{ve}(i-j, j)] \quad (18)$$

### III. RESULTS AND DISCUSSIONS

Asphalt pavement in Vietnam comprises in general two or three asphalt layers, unbound granular material (UGM) layers, and the road base. The total thickness of asphalt layers in general ranges from 12 to 15 cm for national highway and 20 to 25 cm for expressway. In this study, we will analyze the behavior of two-asphalt layer pavement in which conventional (AC) and high modulus asphalt concrete (HMAC) are used. According to our previous study [26], HMAC should be placed in the surface course to maximize the rutting resistance of the whole pavement. Therefore, we will investigate three pavement structures denoted as PS-1 (AC + AC), PS-2 (HMAC + AC), and PS-3 (HMAC + HMAC). The thickness of the asphalt concrete of surface layer (ACS) and asphalt concrete of base layer (ACB) is 5 cm and 7 cm, respectively, and that of UGM layer is 40 cm.

The viscoelastic behavior of HMAC and AC were characterized and predicted using Huet-Sayegh model in Section 3. The UGM and the road base are assumed to be elastic material of 120 MPa of stiffness modulus. The Poisson ratio is assumed to be 0.35 for all layers.

The applied load is a uniformly distributed force moving along the  $x$ -axis representing the application of one-half of a 10-ton axle load as indicated in Fig. 3. The contact pressure is assumed to be 0.6 MPa and the equivalent radius of the tire-print is 0.33 m. The velocity  $v$  of the moving load is 5 km/h representing the low speed traffic areas or 80 km/h for high speed traffic areas. The moving direction of axle load in all cases is from left to right.

In order to verify the calculation results considering the effect of moving load, the calculation results due to a stationary load with equivalent loading time are also shown. The equivalent loading time is:

$$t_L = \frac{2a}{v} \quad (19)$$

where  $2a=0.33$  m is the length (diameter) of the tire-print. The contact pressure  $p(t)$  is assumed to be a constant during the loading time:

$$p(t) = \begin{cases} 0.6 \text{ MPa}, & \text{for } t \in [0, t_L] \\ 0, & \text{for } t > t_L \end{cases} \quad (16)$$

The results were obtained with a homemade program, in which functions for solving multi-layer elastic half-space were adopted from Ref. [27]. The analysis results including the displacement at the surface (deflection), the stress and strain at the surface and at the bottom of ACB layer are presented in next Subsections.

#### A. Displacement at the surface of pavement

Figs. 5-7 depict the deflection due to a moving load at a speed of 5 and 80 km/h. It can be observed that a lower velocity will give rises to a higher value of deflection. The

reason is that the loading time, or the time that the load applies to the structure, is longer with a lower velocity, giving more time for the material to flow. In addition, the effects of HMAC can be clearly observed. Because of higher stiffness, pavements comprising HMAC show smaller deflection than does conventional pavement—namely, the maximal deflection at 5 km/h of PS-3 is 44.1  $\mu\text{m}$ , that of PS-2 is 52.7  $\mu\text{m}$ , and that of PS-1 is 59.5  $\mu\text{m}$ ; the maximal deflection at 80 km/h of PS-3 is 37.4  $\mu\text{m}$ , that of PS-2 is 52.7  $\mu\text{m}$ , and that of PS-1 is 59.5  $\mu\text{m}$ .

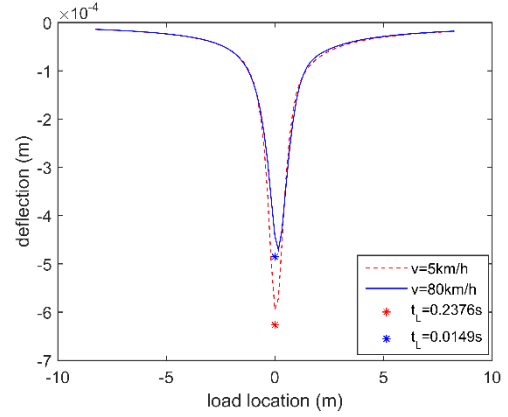


Fig. 5. Deflection at the surface of PS-1 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

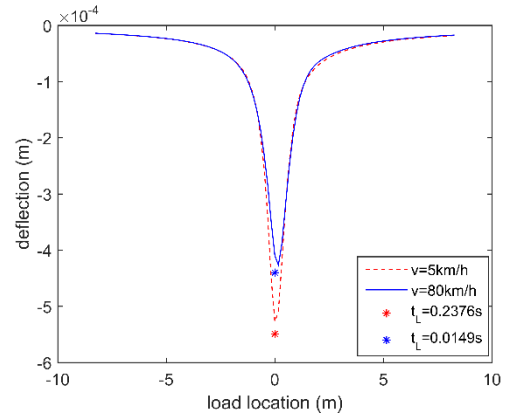


Fig. 6. Deflection at the surface of PS-2 due to moving load at speed of 5 and 80 km/h and due to stationary load with equivalent loading time

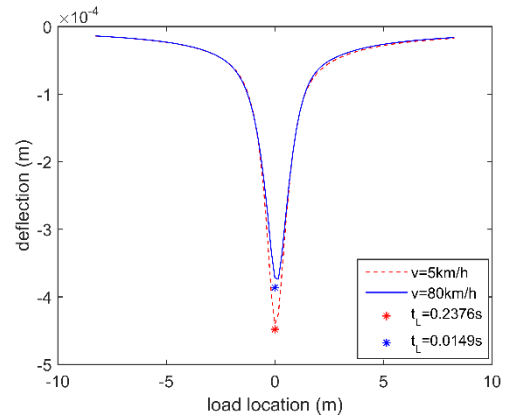


Fig. 7. Deflection at the surface of PS-3 due to moving load at speed of 5 and 80 km/h and due to stationary load with equivalent loading time

The deflections of pavement due to a stationary load with equivalent loading time are also presented in Figs. 5-7, which

show that the application of a stationary load with equivalent loading time results in a slightly higher deflection compared to that of a moving load. This can be explained by the fact that the contact pressure is assumed to be constant during the loading time. However, the different between these values is small, less than 5.4% in all cases as can be seen in the figures.

In practice, the contact pressure is normally assumed to have a haversine wave shape in the loading time:

$$p(t) = \begin{cases} 0.6 \times \sin^2(\pi t / T_L) \text{ MPa, for } t \in [0, t_L] \\ 0, \text{ for } t > t_L \end{cases} \quad (17)$$

The calculation with this assumption of contact pressure underestimates the deflection of pavement. In effect, the deflections at 5 km/h and 80 km/h of PS-1, PS-2 and PS-3 are 54.5, 48.5 and 41.0  $\mu\text{m}$ , and 43.3, 40.1 and 36.5  $\mu\text{m}$ , which are approximately 8.4% smaller than that calculated with moving loading. These results are not as good as those obtained with a stationary load with constant contact pressure. Therefore, in next Subsections, we will not present the results obtained with stationary load with constant contact pressure.

### B. Stress and strain at the top of ACS layer

Figs. 8-10 show the stress at the surface of the pavement due to moving loading. The normal stress ( $\sigma_{xx}, \sigma_{yy}$ ) is negative (compressive) when the axle load is located at the observation point ( $x=0$ ). Interestingly, the stress is positive (tensile) when the axle load is located near this point. The tensile stress is highest when the moving load just overpasses the observation point, at position  $x=5a$ . It can also be seen that  $\sigma_{xx}$  at the surface of the pavement PS-1 is the highest followed by that of PS-2 and PS-3 (see Tab. 1). The strains at the surface of pavement are also presented in Figs. 13-15. Similar to the normal stresses, the normal strains ( $\varepsilon_{xx}, \varepsilon_{yy}$ ) is negative (compressive) when the axle load is located at the observation point and positive (tensile) when the axle load is located near this point. The tensile strain is highest when the moving load is just approaching the observation point, at  $x=-3a$ . Similar to the normal stress, the maximal normal strain  $\varepsilon_{xx}$  at the pavement surface of PS-1 is the highest followed by that of PS-2 and PS-3 (see Tab. 1). Such tensile stress and strain occurring at the surface of the pavement will initiate the top-down cracking distress. Based on the obtained results, HMAC is found to reduce the maximal value of tensile stress and strain at the surface of pavement. Therefore, application of HMAC is expected to increase the resistance to top-down cracking of the pavement.

In order to emphasize the effect of moving loading on the stress and strain at pavement surface, the distribution of stress and strain at pavement surface caused by a stationary load with equivalent loading time was also computed and shown in Figs. 11 and 12 and summarized in Tab.12. The maximal value of  $\sigma_{xx}$  at the surface of PS-1 is 0.25 MPa at 5 km/h while the corresponding maximal value of  $\varepsilon_{xx}$  is 99.7  $\mu\text{e}$ . Compared to the results calculated with moving load, the maximal value of stress calculated with stationary load of constant contact pressure is smaller while the maximal value of strain is larger. This can be explained by the fact that the effects of applied forces are accumulated along the moving path and the

effective loading time due to a moving load is smaller than that due to a stationary load. This observation is valid for all circulation speeds and types of pavement considered. Thus, the calculation based on stationary loading underestimates the tensile stress, yet overestimates the tensile strain at the pavement surface.

TABLE II. STRESS AND STRAIN AT PAVEMENT SURFACE

Type of loading	Maximal stress or strain	Pavement		
		PS-1	PS-2	PS-3
Moving load	$\sigma_{xx}$ (MPa)	0.56 (0.78)	0.67 (0.53)	0.44 (0.44)
	$\sigma_{yy}$ (MPa)	0.25(0.34)	0.22 (0.17)	0.18 (0.14)
	$\varepsilon_{xx}$ ( $\mu\text{e}$ )	74.8 (31.4)	39.8 (22.2)	31.9 (17.1)
	$\varepsilon_{yy}$ ( $\mu\text{e}$ )	-0.2 (-0.1)	- 0.2 (-0.1)	-0.1 (-0.1)
Stationary load of constant contact pressure	$\sigma_{xx}$ (MPa)	0.25 (0.25)	0.41 (0.29)	0.22 (0.23)
	$\sigma_{yy}$ (MPa)	~ 0	~ 0	~ 0
	$\varepsilon_{xx}$ ( $\mu\text{e}$ )	99.7 (47.1)	51.6 (28.0)	35.7 (20.0)
	$\varepsilon_{yy}$ ( $\mu\text{e}$ )	~ 0	~ 0	~ 0

<sup>a</sup>. The values in parentheses are obtained at velocity of 80km/h

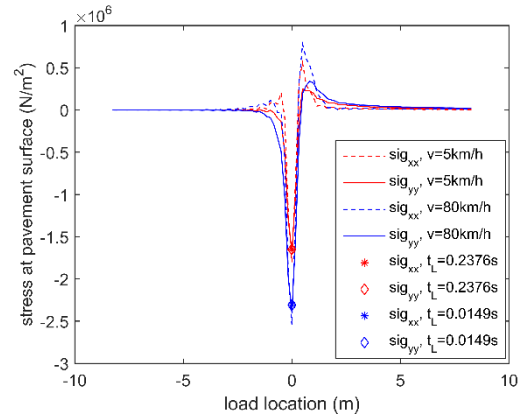


Fig. 8. Stress at the surface of PS-1 due to moving load at speed of 5 and 80 km/h and due to stationary load with equivalent loading time

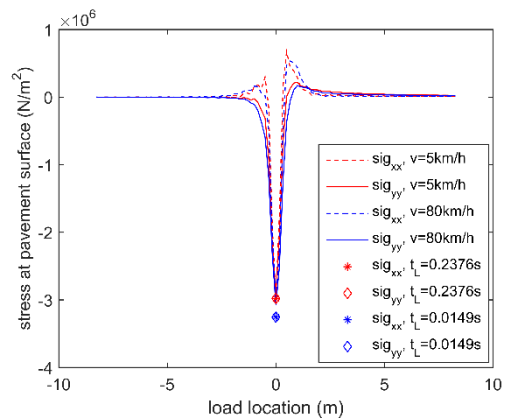


Fig. 9. Stress at the surface of PS-2 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

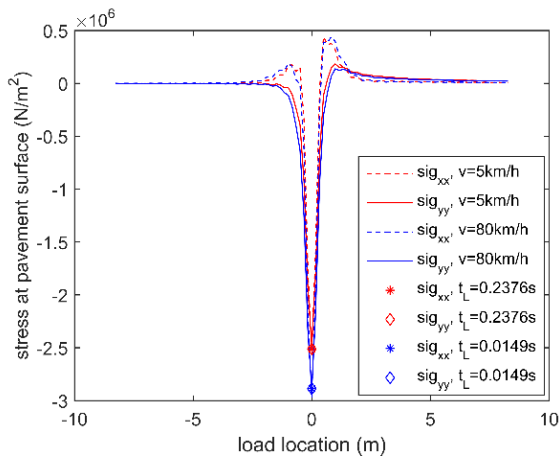


Fig. 10. Stress at the surface of PS-3 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

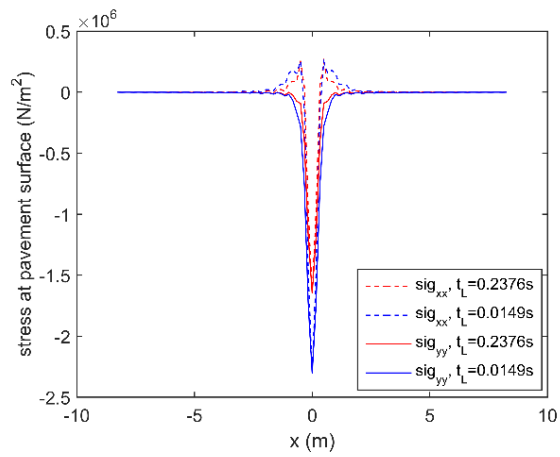


Fig. 11. Distribution of stress at pavement surface of PS-1 due to stationary load with loading time corresponding to speed of 5 and 80 km/h

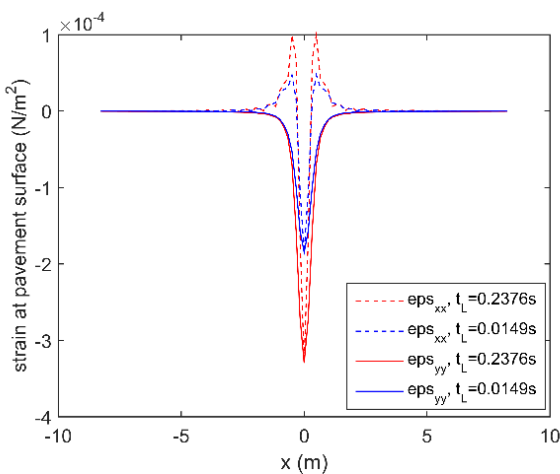


Fig. 12. Distribution of stress at pavement surface of PS-1 due to stationary load with loading time corresponding to speed of 5 and 80 km/h

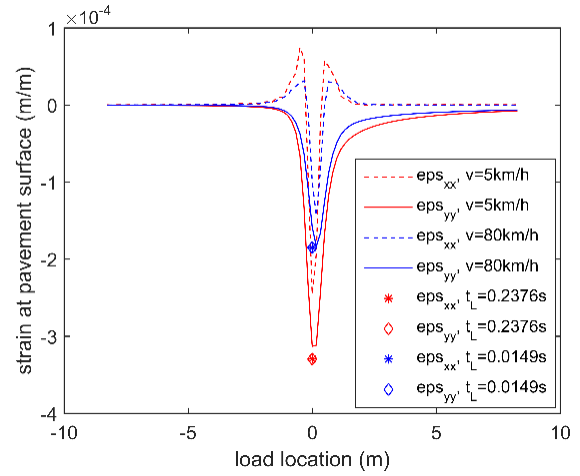


Fig. 13. Strain at the surface of PS-1 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

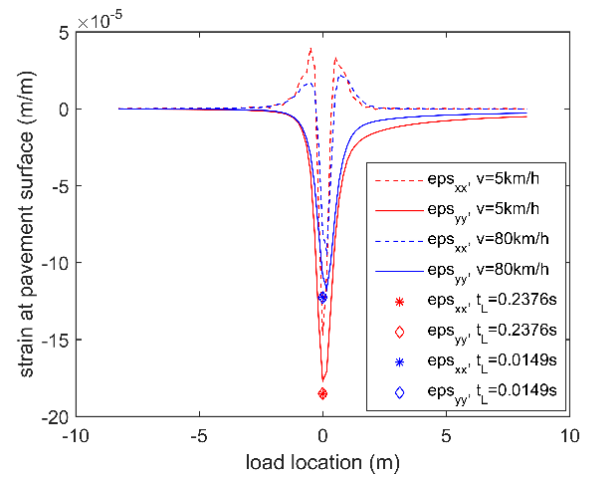


Fig. 14. Strain at the surface of PS-2 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

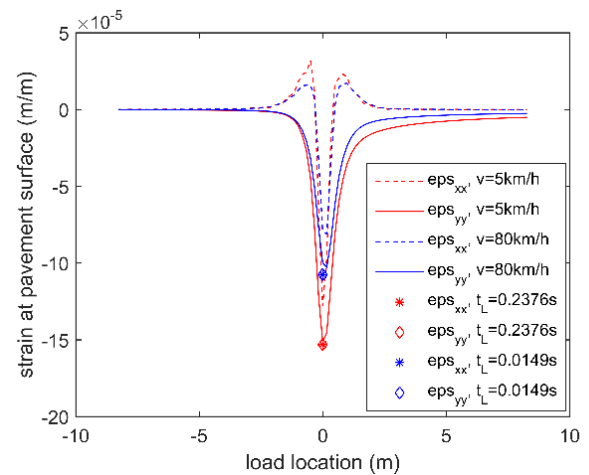


Fig. 15. Strain at the surface of PS-3 due to moving load at speed of 5 and 80 km/h and due to stationary load with equivalent loading time

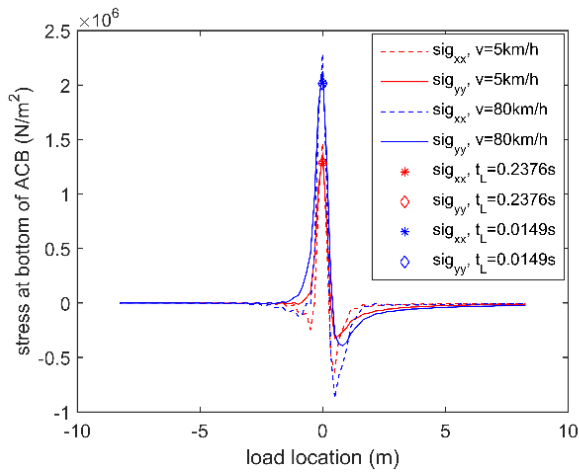


Fig. 16. Stress at the bottom of ACB of PS-1 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

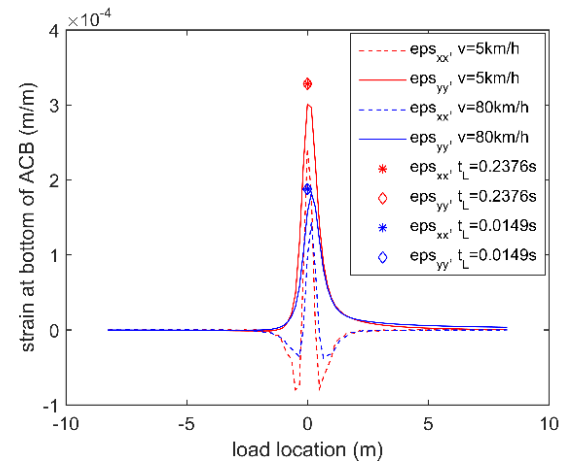


Fig. 19. Strain at the bottom of ACB of PS-1 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

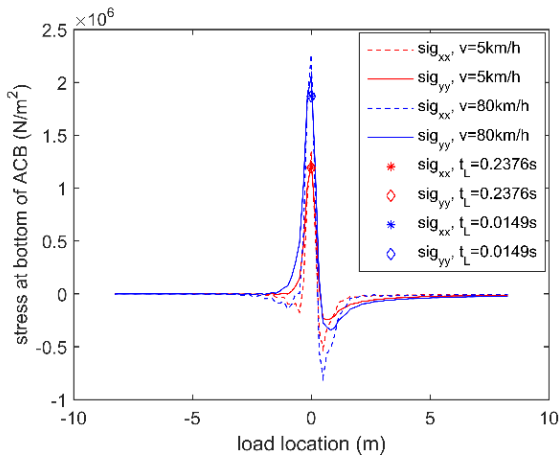


Fig. 17. Stress at the bottom of ACB of PS-2 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

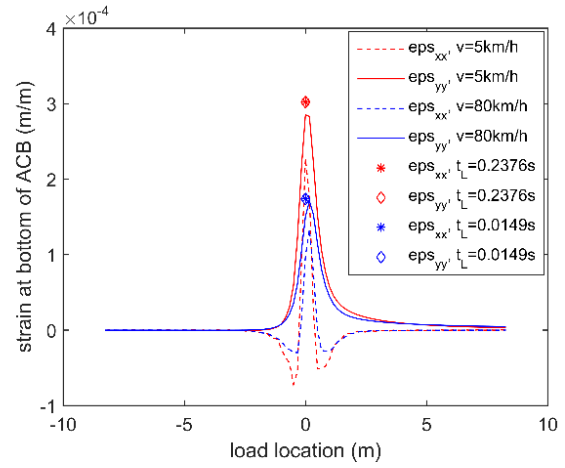


Fig. 20. Strain at the bottom of ACB of PS-2 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

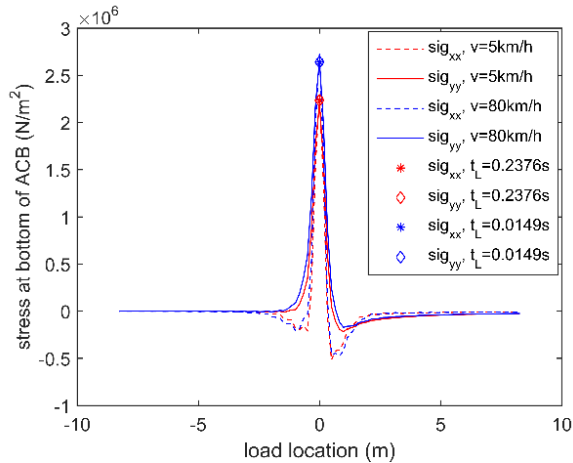


Fig. 18. Stress at the bottom of ACB of PS-3 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times

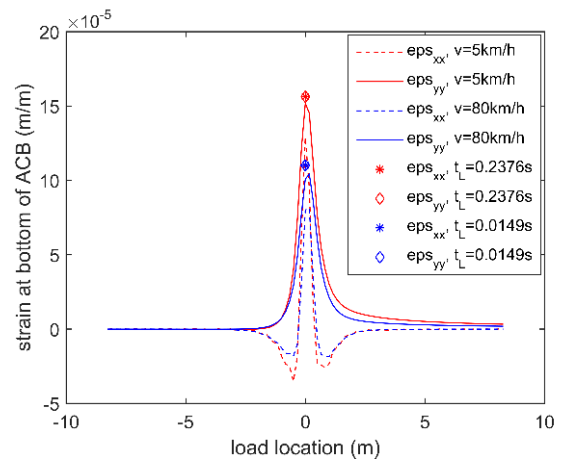


Fig. 21. Strain at the bottom of ACB of PS-3 due to moving load at speed of 5 and 80 km/h and due to stationary loads with equivalent loading times



### C. Stress and strain at the bottom of ACB layer

The stress and strain at the bottom of ACB layer caused by moving loading are shown in Figs. 16-21 and summarized in Tab. 3. It can be seen that the tensile stress and strain are maximal when the moving load is right at the observation point or just overpasses the observation points ( $x = a$ ). In all cases, the higher the velocity is the higher the maximal value of tensile stress at the bottom will be because the material is stiffer at shorter loading time as previously discussed. Similarly, the calculation based on stationary loading underestimates the tensile stress yet overestimates the tensile strain at the bottom of ACB layer.

TABLE III. STRESS AND STRAIN AT BOTTOM OF ASPHALT LAYER

Type of loading	Maximal stress or strain	Pavement		
		PS-1	PS-2	PS-3
Moving load	$\sigma_{xx}$ (MPa)	1.47 (2.29)	1.35 (2.26)	2.29 (2.74)
	$\sigma_{yy}$ (MPa)	1.34 (2.13)	1.23 (2.06)	2.21 (2.66)
	$\varepsilon_{xx}$ ( $\mu\epsilon$ )	242 (144)	227 (133)	129 (114)
	$\varepsilon_{yy}$ ( $\mu\epsilon$ )	301 (182)	285 (171)	152 (104)
Stationary load of constant contact pressure	$\sigma_{xx}$ (MPa)	1.28 (2.01)	1.19 (1.87)	2.24 (2.64)
	$\sigma_{yy}$ (MPa)	1.28 (2.01)	1.19 (1.87)	2.24 (2.64)
	$\varepsilon_{xx}$ ( $\mu\epsilon$ )	328 (188)	302 (174)	156 (110)
	$\varepsilon_{yy}$ ( $\mu\epsilon$ )	328 (188)	302 (174)	156 (110)

<sup>b</sup>. The values in parentheses are obtained at velocity of 80km/h

### IV. CONCLUSIONS

The analysis of LVE response of flexible pavements comprised of conventional asphalt concrete and high modulus asphalt concrete subjected to moving loading has been successfully performed based on an analytical approach. Through a variety of numerical results in this study, the following conclusions could be drawn:

- The deflections of flexible pavement obtained with a moving load are comparable with those obtained with a stationary load. The difference between the obtained deflections is less than 5.4% when the contact pressure is assumed constant during the during loading time and less than 8.4% when haversine waveshape of contact pressured is used.
- The maximal stress at the observation point caused by a moving load is higher than that caused by a stationary load due to the accumulative effects of applied forces on the moving path. The maximal strain caused by a moving load is lower than that caused by a stationary load because the effective loading time due to a moving load is smaller than that due to a stationary load. In other words, the calculation based on stationary loading underestimates the tensile stress, yet overestimates the tensile strain.
- Using HMAC in the ACS layer leads to a reduction in maximal stress and strain at the surface of pavements and at the bottom of ACB layer. Therefore, HMAC ACS layers is expected to increase the resistance to top-down and bottom-up cracking of the pavement.
- Using HMAC in both the ACS and ACB layers of a two-asphalt layer pavement results in an increase in stress and a decrease in strain at the bottom of ACB layer. As a result, advanced fatigue characterization and prediction methods should be considered to evaluate the effects of two HMAC layers on the fatigue life of the pavement

### REFERENCES

- [1] D.E. Newcomb, K.R. Hansen, "Mix Type Selection for Perpetual Pavements", in Proceedings of International Conference on Perpetual Pavements, Ohio University, Ohio, 2006.
- [2] S. Islam et al., "Mechanistic-Empirical design of perpetual pavement", Road Mater Pavement, vol. 21, pp. 1224-1237, 2020.
- [3] H.T.T. Nguyen, N.H. Nguyen, "Using a Non-local Elastic Damage Model to Predict the Fatigue Life of Asphalt Pavement Structure", in Proceedings of the International Conference on Advances in Computational Mechanics 2017, Singapore: Springer, 2018, pp. 47-63.
- [4] H.T.T. Nguyen, V.T. Tran, "Analysing the interlayer shear stress of asphalt pavement composed of conventional and high modulus asphalt", Journal of Science and Technology in Civil Engineering, vol. 13, pp. 85-92, 2019.
- [5] H.T.T. Nguyen, "Modelling the mechanical behaviour of asphalt concrete using the Perzyna viscoplastic theory and Drucker-Prager yield surface", Road Mater Pavement, vol. 18, pp. 264-280, 2017.
- [6] Decision 858-MOT, "Applying current technical standards to strengthen the quality of management of design and construction of hot mix asphalt pavement of high traffic roadway (in Vietnamese)", Hanoi: Ministry of Transport, 2014.
- [7] H.T.T. Nguyen, T.N. Tran, "Effects of crumb rubber content and curing time on the properties of asphalt concrete and stone mastic asphalt using dry process", International Journal of Pavement Research and Technology, vol. 13, pp. 238-244, 2018.
- [8] T.N. Tran et al., "Semi-flexible Material: The Sustainable Alternative for the Use of Conventional Road Materials in Heavy-Duty Pavement", in Congrès International de Géotechnique – Ouvrages – Structures, Singapore: Springer, 2018, pp. 17-24.
- [9] Tien-Tho Do et al., "Effects of Forta-Fi fiber on the resistance to fatigue of conventional asphalt mixtures", submitted to 5th International Conference on Green Technology and Sustainable Development which will be held in 2020.
- [10] G.D. Airey, B. Rahimzadeh, A. Collop, "Evaluation of the linear and non-linear viscoelastic behaviour of bituminous binders and asphalt mixtures", In Proceedings of 6th International Conference on the Bearing Capacity of Roads, Rotterdam: A.A. Balkema Publishers, 2002, pp. 799-812.
- [11] H. Di Benedetto et al., "Linear viscoelastic behaviour of bituminous materials: From binders to mixes", Road Mater Pavement, vol. S1, pp. 163-202, 2004.
- [12] R. Bonaquist, D.W. Christensen, "Practical procedure for developing dynamic modulus master curves for pavement structural design", Transportation Research Record: Journal of the Transportation Research Board, 1929, pp. 208-217, 2005.
- [13] J. Judycki, "A new viscoelastic method of calculation of low-temperature thermal stresses in asphalt layers of pavements", Int J Pavement Eng, vol. 19, pp. 24-36, 2016.
- [14] 22 TCN 211-06, "Flexible pavement – Requirements and specifications for design", Hanoi: Ministry of Transport, 2006.
- [15] Q.T. Nguyen et al., "Nonlinearity of bituminous materials for small amplitude cyclic loadings", Road Mater Pavement, vol. 20, pp. 1571-1585, 2019.
- [16] H.T.T. Nguyen, V.T. Tran, "Analysis of stress and strain in flexible pavement structures comprised of conventional and high modulus asphalt using viscoelastic theory", in Proceeding of Rebuilt2019 Conference, Iasi, 2019, in press.
- [17] V.P. Bui, Q.T. Nguyen, "Dynamic tests for determining mechanical properties of asphalt binders 60/70, 35/50 and PMB3 using DSR and metravib DMA equipment", The Transport and Communications Science Journal, vol. 71, pp. 583-594, 2020.
- [18] H.T.T. Nguyen, D.L. Nguyen, V.T. Tran, M.L. Nguyen, "Finite element implementation of Huet-Sayegh and 2S2P1D models for

- analysis of asphalt pavement structures in time domain”, Road Mater Pavement, 2020. doi: 10.1080/14680629.2020.1809501.
- [19] E. Levenberg, “Viscoelastic Pavement Modeling with a Spreadsheet”, in Proceedings of the Eighth Intl. Conf. on Maintenance and Rehabilitation of Pavements, Singapore: Research Publishing, 2016, pp. 746-755.
- [20] K. Chatti et al., “Enhanced Analysis of Falling Weight Deflectometer Data for Use With Mechanistic-Empirical Flexible Pavement Design and Analysis and Recommendations for Improvements to Falling Weight Deflectometers. Report No. FHWA-HRT-15-063”, Georgetown Pike: U.S. Department of Transportation, 2017.
- [21] C. Huet, “Etude par une méthode d’impédance du comportement visco-élastique des matériaux hydrocarbures (Ph.D. Thesis)”, Paris: Faculté des Sciences de Paris, 1963.
- [22] G. Sayegh, “Contribution à l’étude des propriétés visco-élastiques des bitumes purs et des bétons bitumineux (Ph.D. Thesis)”. Paris: Sorbonne University, France, 1965.
- [23] F. Olard, H. Di Benedetto, “General “2S2P1D” model and relation between the linear viscoelastic behaviours of bituminous binders and mixes”, Road Mater and Pavement Design, vol. 4, pp. 185-224, 2003.
- [24] R.A. Schapery, S.W. Park, “Methods of interconversion between linear viscoelastic material functions. Part II—An approximate analytical method”, International Journal of Solids and Structures, vol. 36, pp. 1677-1699, 1999.
- [25] M.K. Charyulu, “Theoretical stress distribution in an elastic multi-layered medium (Ph.D. Thesis)”. Ames: Iowa State University of Science and Technology, 1964.
- [26] T.N. Phan, H.T.T. Nguyen, “Evaluating the possible use of high modulus asphalt mixtures in flexible pavements in Vietnam”, in Proceedings of The International Conference on Sustainable Civil Engineering and Architecture (ICSCEA) 2019, Ho Chi Minh City, 2019.
- [27] A. Al-Rumaihi, “Multi-layer Elastic Analysis”, retrieved from <https://www.mathworks.com/matlabcentral/fileexchange/69465-multi-layer-elastic-analysis>.

# Effects of Forta-fi Fiber on the Resistance to Fatigue of Conventional Asphalt Mixtures

Tien-Tho Do

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thodt@hcmute.edu.vn

Duy-Liem Nguyen

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
liemnd@hcmute.edu.vn

Vu-Tu Tran

Department of Transport Engineering,  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tutv@hcmute.edu.vn

H.T. Tai Nguyen

Department of Transport Engineering, Faculty of Civil Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
tainht@hcmute.edu.vn

**Abstract**— In order to deal with the issues of rutting in Vietnam, asphalt mixtures have been highly modified using a variety of additives, e.g. SBS, Polyethylene, crumb rubber or even produced with less bitumen so as to improve its resistance to permanent deformation. As a result, the flexibility of modified asphalt mixtures much decreases compared to unmodified mixtures, giving rise to a reduction in the resistance to fatigue cracking. Moreover, the thickness of flexible pavement structures in Vietnam is typically small ranging from 12 to 15 cm in asphalt layers while overloaded vehicles cannot be effectively controlled. Consequently, fatigue cracking occurs more and more often and early after opening to traffic. This paper aims at evaluating the effectiveness of Forta-fi fiber in improving the resistance to fatigue of conventional asphalt mixtures produced with low bitumen content and a very coarse dense gradation. Marshall method was used for sample compaction and mix design, whereas cyclic indirect tensile test was used for fatigue characterization. The experimental results show that adding 0.1% of Forta-fi fiber to asphalt mixtures significantly improves its resistance to fatigue, up to 2 times higher.

**Keywords**—Forta-fi fiber, asphalt mixtures, resistance to fatigue, flexible pavement component

## I. INTRODUCTION

Rutting and fatigue are two common distresses of asphalt pavement in Vietnam. Rutting normally occurs early after opening to traffic while fatigue cracks take longer time to initiate and propagate to the pavement surface. Once rutting occurs, the pavement surface loses its evenness and smoothness, giving rise to a shortage of traffic safety. As a result, rutting distress attracts much more attention of local authorities and researchers than fatigue cracking does. Many solutions have been examined and applied in the country to improve the resistance to rutting of asphalt pavement, e.g. using polymer modified binder, crumb rubber modified mixture, SMA gradation to maximize the coarse aggregate interlock, semiflexible pavement or even using a very coarse gradation curve to reduce the binder content used in the mixture [1,2,3].

Using little binder in the mixture can improve the resistance to rutting of asphalt pavement in the short-term. However, it reduces the fatigue life of asphalt pavement in

long-term. In effect, based on our observation in the field, fatigue cracks occur more often and sooner than before. Fig. 1 shows fatigue cracks and pothole in the asphalt pavement of the National Highway 1A located near Thu Duc district, Ho Chi Minh City, Vietnam. The figure was taken just 2 years after the opening to service of the road. Therefore, along with controlling the resistance to rutting of asphalt mixture as required in national codes and prediction of service life of the asphalt pavement based on rutting depth as proposed in Ref. [4], the resistance to fatigue of asphalt mixtures and pavement should be examined and predicted before acceptance of the material.



Fig. 1. Fatigue cracks and potholes in National Highway 1A, located near Thu Duc District, Ho Chi Minh City, Vietnam

Fatigue cracking has not been well studied in Vietnam due to the lack of testing devices. In addition, methods for preventing fatigue cracking are not specified in the actual code for design of asphalt pavement [5]. As a result, fatigue cracking has become one of the two most common distresses of pavement structures in the country.

A lot of solutions have been proposed in the literature to improve the resistance to fatigue of asphalt mixtures and pavements, e.g. controlling the thickness of bitumen film in the mixture [6], using high bitumen content like SMA mixture [7], using geogrid to reinforce the resistance to tensile stress of pavement layer [8], using fiber in the mixture composition [9,10]. However, a limited number of solutions have been investigated by local researchers and practitioners for prediction and improvement of resistance to fatigue of asphalt

mixture and pavement, e.g. Refs. [11,12,13]. This study is an attempt to study the resistance to fatigue of a very coarse (low-bitumen) dense gradation asphalt mixture produced with the commonly used 60/70 pen bitumen in Vietnam and its improvement by adding Forta-fi fiber.

## II. MATERIALS AND METHODS

### A. Materials

This experimental study investigates the IDT fatigue behavior of asphalt mixtures. These materials are fabricated with 60/70 pen bitumen using the coarse gradation specified in [3] with nominal maximum particle size of 12.5 mm. The gradation curve studied here is the same as that used in [14] and shown in Fig. 1.

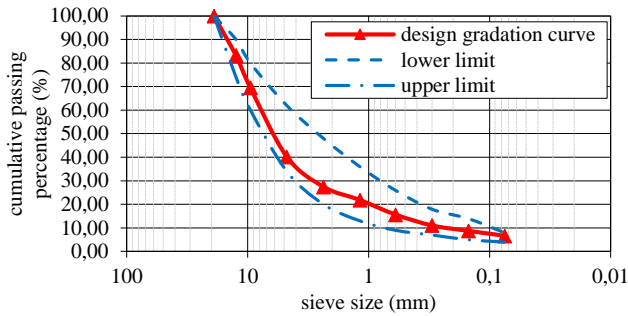


Fig. 2. Design gradation curve used for preparing asphalt samples

BMT Construction Investment J.S. Company has provided all aggregate, filler, and bitumen for this research. The physical properties of these components are the same as those reported in [1].

In order to improve the resistance to fatigue of asphalt mixtures, high tensile strength synthetic fiber provided by Forta-fi Corporation (hereafter called Forta-fi fiber for the sake of simplicity) was added into the composition of asphalt mixtures at a rate of 0.1% of the total mix weight. Forta-fi fiber is a blend of para-aramid and polyolefin fiber (see Fig. 2) whose physical properties were summarized in Table 1.

TABLE I. FUNDAMENTAL PROPERTIES OF FORTA-FI FIBER

Properties	Fiber	
	Para-Aramid	Polyolefin
Specific gravity	1.44	0.91
Tensile strength (MPa)	2.758	N/A <sup>a</sup>
Length (mm)	19	19
Operating temperature (°C)	-73 ÷ 427	N/A <sup>a</sup>

<sup>a</sup> Fiber will melt or plastically deformed during production of asphalt mixture



Fig. 3. Figure of Forta-fi fiber at room temperature and loose condition

### B. Methods

#### 1) Sample preparation

The design of asphalt mixtures was based on Marshall method as specified in Ref. [3]. Cylindrical specimens of asphalt mixtures were prepared using a Marshall compactor such that the final dimensions of sample are 101mm in diameter and 63.5 mm in height. At least 5 × 3 samples were prepared at 5 levels of binder content around its predicted optimal value. This optimum can be determined based on a range of properties varying with binder content such as bulk specific gravity, maximal specific gravity, air void, void in aggregate, void filled with bitumen, Marshall stability and flow. Recently, asphalt mixtures are designed aiming at high percentage of coarse aggregates and low content of bitumen so as to increase the resistance to rutting. Therefore, the optimum binder content used in this study is only 4.8% of total mix weight. The fundamental properties of these mixtures at optimal binder content are summarized in Table II. In order to evaluate the effect of the new component to the mixtures, the experiment is carried out in both cases: samples with and without Forta-fi fibers added. These samples are hereafter coded as D8-F which stands for fiber added samples and D8-N which stands for control samples without fiber added.

TABLE II. FUNDAMENTAL PROPERTIES OF ASPHALT MIXTURES

Properties	Mixture	
	D8-N	D8-F
Maximum specific gravity	2.52	2.52
Bulk specific gravity	2.35	2.36
Air void (%)	6.62	6.46
Void in aggregate (%)	16.63	16.48
Marshall stability (kN)	10.5	12.0
Marshall flow (mm)	2.9	2.8
Indirect tensile strength at 25°C (IDTS) (MPa)	1.09	1.49

#### 2) Fatigue characterisation method

With the aim of evaluating the resistance to fatigue of asphalt mixtures, the Pavetest hydraulic Universal Testing Machine (Fig. 3a) equipped at Ho Chi Minh City University of Technology and Education (HCMUTE) was used, which was implemented with test method EN 12697-24, Annex E (IDT test on cylindrical shaped specimens). The test method can be summarised as follows.

A vertical force of continuous haversine function with amplitude  $P_{cyclic}$  was applied to the sample along its diametral axis as illustrated in Fig. 3b. As a result, the tensile stress also has form of haversine function. For points located at the symmetry line and far enough from the applied force, the tensile stress is maximal and can be calculated by:

$$\sigma_{cyclic} = \frac{2P_{cyclic}}{\pi Dt} \quad (1)$$

where  $D$  and  $t$  is the diameter and thickness of sample. The horizontal displacement of two extremity points (point A and B in Fig. 3b) also has form of haversine function, which is captured by two LVDTs. Assuming that the Poisson ratio of asphalt mixtures is known a priori, the indirect tensile (IDT) stiffness of the sample can be calculated by:

$$S_{IDT} = P_{cyclic} \times \frac{(\nu + 0.27)}{D \cdot \Delta H_{cyclic}} \quad (2)$$

where  $\Delta H_{cyclic}$  is the amplitude of the relative displacement of points A and B in the horizontal direction. Since asphalt mixture shows viscoelastic behavior, the indirect tensile stiffness depends on the temperature and frequency of the applied load. With the intention of imitating the climatic of traffic conditions in Vietnam, the examined temperature and frequency is 25°C and 10Hz as proposed by Tran et al. and Nguyen et al. [12,15].

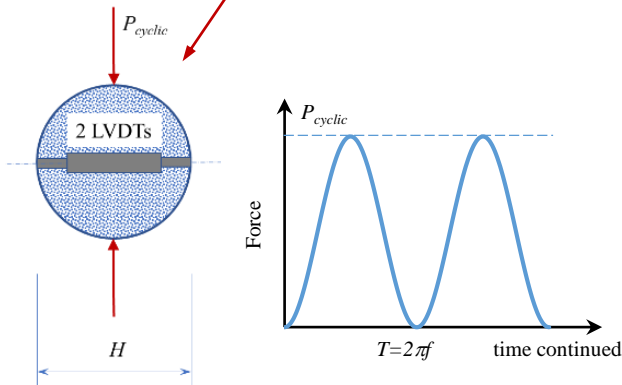


Fig. 4. (a) Figure of 30kN hydraulic UTM at HCMUTE. (b) Illustration of applied force

Under the effect of repeated loading, microcracks in material develop; and therefore IDT stiffness decreases according to the increase of the amount of load cycles. The number of cycles at which IDT stiffness reduces 50% of its initial value is regarded as the fatigue life of tested specimens, noted  $Nf_{50}$ . The curve of fatigue life with respect to applied stress amplitude or initial strain amplitude reflects the fatigue behavior of the mixture studied and is called the characteristic fatigue line (CFL).

### III. RESULTS AND DISCUSSIONS

For the future results and discussions, the nominal damage of the specimen  $D_{nom}$  is defined as following [15]:

$$D_{nom} = 1 - \frac{S_n}{S_0} \quad (3)$$

where  $S_0, S_n$  are respectively the IDT stiffnesses of the sample at initial state (cycle 100) and at cycle  $n > 100$ .

For the tests performed at low stress level, the value of the fatigue life of the samples can reach several million of cycles. To reduce the time of experiment, the tests were stopped when the slope of the nominal damage reaches a stable value or when the number of cycles exceeds 10 000. The fatigue life of the specimen was predicted based on linear regression method:

$$D_{nom}(n) = D_0 + k \cdot n$$

$$Nf_{50} = \frac{0.5 - D_0}{k} \quad (4)$$

where  $k, D_0$  is regression parameters and  $n$  is the number of load cycle.

Five levels of stress magnitudes ( $\sigma_{cyclic}$ ) were investigated all along the study of the fatigue behavior of asphalt mixtures. During the test, a low constant stress ( $\sigma_{const.}$ ) is maintained to keep a positive contact between the specimen and the loading device. Table III summarizes the applied stress magnitudes ( $\sigma_{cyclic}$ ) and constant stresses ( $\sigma_{const.}$ ) for the test on the mixture samples D8-N (without Forta-fi added) and D8-F (with Forta-fi added). The total stress including the static and cyclic one applied to the sample is  $\sigma_{cyclic} + \sigma_{const.}$ . Only one replicate of test is performed in this study due to its high testing cost. The obtained results will be presented and discussed in next paragraphs.

TABLE III. PRESCRIBED STRESS USED IN THIS STUDY

Mixture	Constant stress (kPa)	Cyclic stress (kPa)
D8-N	10	170
	13	243
	19	365
	21	488
	28	733
D8-F	11	170
	12	244
	16	367
	19	488
	27	732

#### A. Effect of Forta-fi fiber on fundamental properties of asphalt mixtures

It can be observed in Table IV that Forta-fi fiber shows its efficiency while improving the fundamental properties of asphalt mixture with a relatively low quantity of fiber added. In fact, the Marshall stability of mixtures added with only 0.1% of Forta-fi (percentage on total mixture weight) increases approximately 12.3% (from 10.5kN to 12kN). On the other hand, this complement fiber permits to ameliorate the IDT strength approximately 36.7% (from 1.09MPa to 1.49MPa). Consequently, it is expected that Forta-fi fiber will improve the resistance to fatigue as well as the resistance to rutting of asphalt mixtures. This work focuses only on the behavior of asphalt mixtures modified with Forta-fi fiber. The effects of fiber on the resistance to rutting of asphalt mixtures will be considered in another work.



### B. Effect of Forta-fi fiber on initial stiffness of asphalt mixture

Fig. 5 shows the initial stiffness of the asphalt mixtures under different applied stresses of the samples with and without Forta-fi added. We observe a significant increase of the initial stiffness of the sample due to the effect of the Forta-fi fiber added into the asphalt mixture. The IDT stiffness of samples modified with Forta-fi fiber is 1.4 times that of unmodified samples in average. Therefore, the initial strain of samples modified with Forta-fi fiber is only 0.7 time that of unmodified samples. As a result, Forta-fi fiber is predicted to extent the fatigue life of asphalt mixtures.

Nonlinearity behavior of asphalt mixtures is also observed in Fig. 5. The larger the stress (or strain) level is applied, the smaller initial IDT stiffness of mixtures is obtained. This observation is in good agreement with that reported in [15,16].

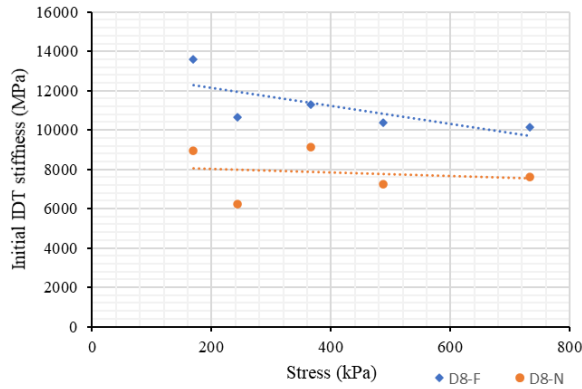


Fig. 5. Effect of Forta-fi fiber on initial stiffness of asphalt mixture

### C. Evolution of nominal damage

Figures 6 and 7 show the evolutions of the nominal damage of D8-N and D8-F samples in function of load cycles. It is clear that a larger level of cyclic stress will give rise to a shorter fatigue life. Due to the oscillation of the captured data during the test and the improper contact between the testing frame and the sample, the value of stiffness at a cycle oscillates around its average value. This phenomenon explains the existence of a small negative value of  $D_{nom}$  at the beginning of the test. The fatigue lives ( $Nf_{50}$ ) of samples are predicted using Eq. (4) for the cases that the number of load cycles exceeds 10,000 or directly obtained from the tests for the cases that the test stops at less than 10,000 load cycles. Table IV presents the fatigue lives of both D8-F and D8-N mixtures samples under different cyclic stresses from 170 kPa to 733 kPa.

In order to better understand the fatigue behavior of mixtures, a non-dimensional quantity defined as the ratio of stress to IDT strength was used to plot the characteristic fatigue line of mixtures (CFL) as can be seen in Fig. 6. The CFL of all mixtures follows well the power law with high value of  $R^2$  coefficient. Interestingly, both D8-F and D8-N mixtures share the same CFL with high value of  $R^2 = 0.989$ :

$$Nf_{50} = 122.04 \times \left( \frac{\sigma_{cyclic}}{IDTS} \right)^{-3.08} \quad (5)$$

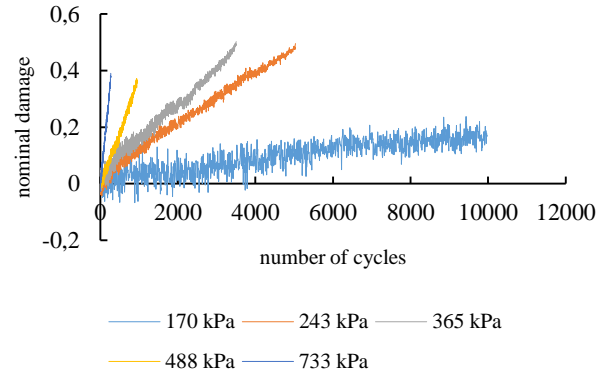


Fig. 6. Evolution of the nominal damage of D8-N samples at different level of initial strain

However, due to higher IDT strength, the D8-F samples have longer fatigue life than that of D8-N samples. In effect, the fatigue life of mixtures modified with Forta-fi fiber is at least 2 times longer than that of unmodified mixtures at any level of stress as presented in the Table IV. The largest difference of the fatigue life between the D8-F and D8-N samples is 2.43 times, at the level of cyclic stress 365 kPa.

TABLE IV. ESTIMATED FATIGUE LIFE OF SAMPLES

Mixture	Cyclic stress (kPa)	Initial strain	Initial stiffness (MPa)	Fatigue life $Nf_{50}$ (cycle)
D8-N	170	38.9	8964.1	28,361
	243	79.3	6215.2	12,755
	365	81.9	9140.3	3,716
	488	137.9	7257.8	1,297
	733	197.8	7610.0	357
D8-F	170	25.6	13587.2	84,558
	244	47.0	10656.9	37,878
	367	65.9	11302.8	12,755
	488	95.8	10393.6	4,064
	732	148.7	10138.2	1119

It should be noted that, in IDT fatigue test, one cannot perform strain control mode because the IDT stiffness is not homogeneously distributed in the sample due to degradation under cyclic load. In order to figure out the fatigue behavior of mixtures in function of strain level, one can plot the fatigue life versus the initial strain level as shown in Fig. 8. In a similar manner, the CFL based on initial strain of both D8-F and D8-N mixtures are almost identical. The CFL for these two mixtures based in initial strain can be linearly regressed with high value of  $R^2 = 0.948$ :

$$Nf_{50} = 6.93 \times 10^8 \times (\epsilon_{initial})^{-2.66} \quad (6)$$

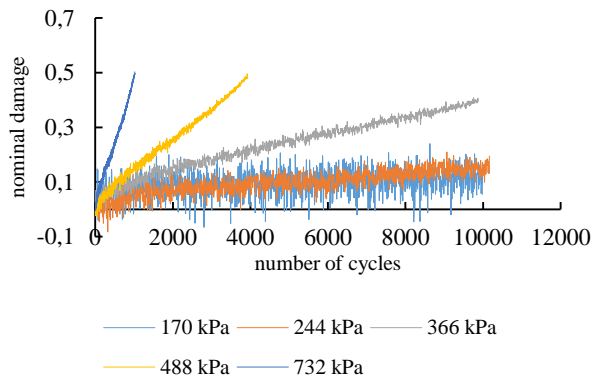


Fig. 7. Evolution of the nominal damage of D8-F samples at different level of initial strain

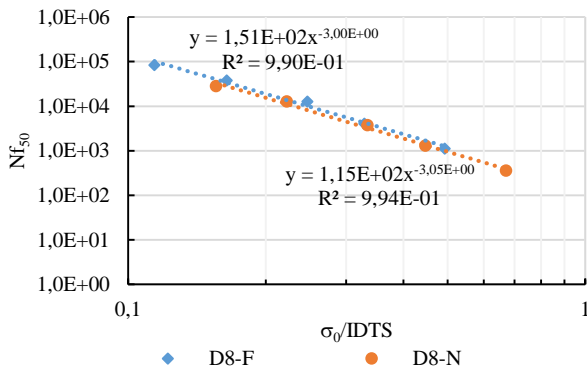


Fig. 8. Fatigue life versus stress level of D8-N and D8-F samples at 25°C, 10Hz

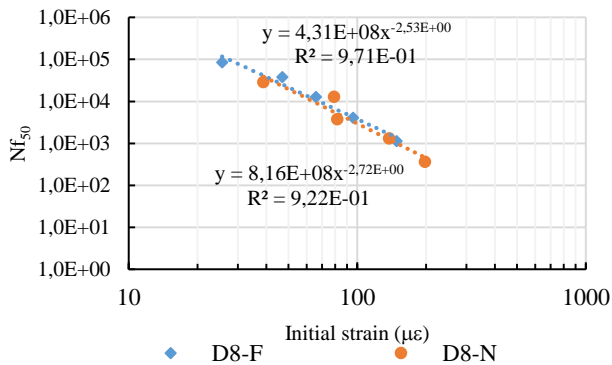


Fig. 9. Fatigue life versus initial strain level of D8-N and D8-F samples at 25°C, 10Hz

Since the initial IDT stiffness of D8-F is larger than that of D8-N, the fatigue life of D8-F is always larger than that of D8-N. Based on the CFL in Eq. (6), the fatigue life of mixtures modified with Forta-fi fiber is at least 2 times that of unmodified mixtures at any level of strain because the initial IDT stiffness of mixtures modified with Forta-fi fiber is higher than that of unmodified mixtures.

#### IV. CONCLUSIONS

The effect of Forta-fi fiber on the resistance to fatigue of low-bitumen conventional asphalt mixtures has been successfully investigated. Through a variety of experimental

results in this study, the following conclusions could be drawn:

- Forta-fi fiber increases the IDT strength (approximately 36.7%) and stiffness of asphalt mixtures (approximately 40%).
- The characteristic fatigue line of both mixtures modified with Forta-fi fiber and unmodified mixtures is almost the same and obeys well the power law with high value of  $R^2$  coefficient.
- The resistance to fatigue of mixture modified with Forta-fi fiber is much higher than that of unmodified mixture—namely, at least 2 times higher.

#### REFERENCES

- [1] H.T.T. Nguyen, T.N. Tran, “Effects of crumb rubber content and curing time on the properties of asphalt concrete and stone mastic asphalt using dry process”, *Int. J. Pavement Res. and Technol.*, vol. 13, pp. 238-244, 2018.
- [2] T.N. Tran, H.T.T. Nguyen, K.S. Nguyen, N.T.H. Nguyen, “Semi-flexible Material: The Sustainable Alternative for the Use of Conventional Road Materials in Heavy-Duty Pavement”, in *Congrès International de Géotechnique – Ouvrages – Structures*, Singapore: Springer, 2018, pp. 17-24.
- [3] Decision 858-MOT, “Applying current technical standards to strengthen the quality of management of design and construction of hot mix asphalt pavement of high traffic roadway (in Vietnamese)”, Hanoi: Ministry of Transport, 2014.
- [4] H.T.T. Nguyen, “Modelling the mechanical behaviour of asphalt concrete using the Perzyna viscoplastic theory and Drucker–Prager yield surface”, *Road Mater Pavement*, vol. 18, pp. 264-280, 2017.
- [5] 22 TCN 211-06, “Flexible pavements – Requirements and specifications for design”, Hanoi: Ministry of Transport, 2006.
- [6] Setra-LCPC, Conception et dimensionnement de structures de chaussée-Guide technique, Service d’études techniques des routes et autoroutes, Paris: Laboratoire central des ponts et chaussées, 1994.
- [7] K. Blaziejowski, *Stone Matrix Asphalt - Theory and Practice*, 1<sup>st</sup> ed., Boca Raton: CRC Press, 2011.
- [8] A. Zofka, M. Maliszewski, D. Maliszewska, “Glass and carbon geogrid reinforcement of asphalt mixtures”, *Road Mater Pavement*, vol. 18, pp. 471-490, 2017.
- [9] R.S. McDaniel, *Fiber Additives in Asphalt Mixtures*, Washington: Transportation Research Board, 2015.
- [10] E. Rohrbough, “Effect of FORTA-FI Fibers on the Rutting Potential, Dynamic Modulus, Flow Number, and Fatigue of Asphalt Concrete Modulus, Flow Number, and Fatigue of Asphalt (Master thesis)”, Morgantown: West Virginia University, 2018.
- [11] H.T.T. Nguyen, N.H. Nguyen, “Using a Non-local Elastic Damage Model to Predict the Fatigue Life of Asphalt Pavement Structure”, in *Proceedings of the International Conference on Advances in Computational Mechanics 2017*, Singapore: Springer, 2018.
- [12] T.L. Tran, V.C. La, X.D. Nguyen, “Experimental study on the resistance to fatigue of asphalt concrete used for pavement surface in Vietnam (in Vietnamese)”, *Journal of Transportation*, vol. 4, 2015.
- [13] T.L. Tran, M.L. Nguyen, “The studying of the resistance to fatigue of asphalt concrete in Vietnam”, *Journal of Transportation (in Vietnamese)*, vol. 6, pp. 23 – 24, 2012.
- [14] H.T.T. Nguyen, V.T. Tran, “Analysis of stress and strain in flexible pavement structures comprised of conventional and high modulus asphalt using viscoelastic theory”, in *Proceeding of Rebuilt2019 Conference*, Iasi, 2019, in press.
- [15] H.T.T. Nguyen, A.T. Le, V.T. Tran, D.L. Nguyen, “Fatigue characterization of conventional and high rutting resistance asphalt mixtures using the cyclic indirect tensile test”, in *Congrès International de Géotechnique – Ouvrages – Structures*, Singapore: Springer, 2018, pp. 579-584.
- [16] H. Di Benedetto, Q.T. Nguyen, C. Sauzéat, “Nonlinearity, Heating, Fatigue and Thixotropy during Cyclic Loading of Asphalt Mixtures”, *Road Mater Pavement*, vol. 12, pp. 129-158, 2011.

# Application of Element Combine39 to Reflect the Nature of Newly Puzzle Shaped Crestbond Rib Shear Connector in Composite Beam

Pham Duc Thien

Department of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
phamducthien@hcmute.edu.vn

Dao Duy Kien\*

Department of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
kiendd@hcmute.edu.vn

Nguyen Van Khoa

Department of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
khoanv@hcmute.edu.vn

Nguyen Thanh Hung

Department of Civil Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
hungnt@hcmute.edu.vn

**Abstract**— In this paper, the three-dimensional finite element model was used to simulate the overall flexural behavior of simple supported composite beams with four-point bending scheme. The investigation was focused on evaluation of load-deflection, relative slip between concrete slab and steel girder, development of strain in concrete slab as well as steel girder. The effect of concrete strength and dimensions of slab were taken into account. The results obtained from modelling are in agreement with results derived from experimental push out tests. Effect of concrete compressive strength and concrete slab width are displayed in detail through deflection and relative slip. There is good agreement between results of relative slip between concrete slab and steel girder obtained from modeling and experimental push out test results. It is supposed that using element COMBIN39 is correct enough to reflect the nature of Perfbond connector.

**Keywords**— Shear connector, Newly puzzle shaped connector, Concrete-steel composite beam, FEM analysis

## I. INTRODUCTION

A newly puzzle shape of crestbond rib connector with a "u" shape was proposed by [1] and application of this type of shear connector on composite structure was also investigated by experiment [2].

A three-dimensional finite element model of composite structure has been developed by [3] using the ABAQUS, the contact behavior between the steel and concrete part, the nonlinear material model curve for steel and the linear material model curve for concrete are assumed in the FEM model. All these studies were focused just on the push-out-test of their corresponding models, but these models were not used to investigate in more detail either the effect of particular structural parameters or other aspects of the system behavior.

To improve the FEM method, (Queiroz et al. 2007, Kraus and Wurzer 1997) has also performed the investigations of three-dimension model using ANSYS, this study focused on

the evaluation of full and partial shear connection in composite beam. The proposed three-dimensional FE model is able to simulate the overall flexural behavior of simply supported composite beams subjected to either concentrated or uniformly distributed loads. This covers: load deflection behavior, longitudinal slip at the steel-concrete interface, distribution of stud shear force and failure modes.

In this study, three-dimensional FEM models of composite beam were developed continuously, in which all the main structural parameters and associated nonlinearities are included (concrete slab, steel beam and shear connectors). Other features such as different shape of concrete slab, concrete strength was considered. Then, the deflection, ultimate load, and strains of the concrete, steel beam, and Perfbond the relative slip between the steel beam and the concrete slab at the end of the beams; and the failure mechanism were observed. The goals of this study include:

- All the main structural parameters and associated nonlinearities are included
- The application of FEM analysis to replace the experiment on the real composite beam based on the fully three-dimension FEM model
- Investigating the overall structural system behavior when different concrete compressive strengths and shape of cross section of concrete slab are used in the slab and in the associated push-out tests.

## II. EASE OF USE FINITE ELEMENT MODEL

### A. Specimens arrangement

To examine the applicability and composite behaviors of the newly puzzle shaped of crestbond shear connector in composite beams under loading, full-size composite beams were prepared (B1). Beam had a 4000 mm of length and a concrete slab that was 120 mm thick and 600 mm wide. Two longitudinal rebars with a diameter of 12 mm were installed in this concrete slab. An opposite T beam with a height of

264 mm and flange width of 200 mm was used. For the shear connector, a continuous crestbond rib shear connector with a height of 70 mm and a thickness of 8 mm was attached to the top flange of the steel H beam, as shown in Fig. 1.

The shear strength of the newly puzzle shaped of crestbond rib shear connector was determined by the push-out test results and suggested design equation [3], and the degree of the shear connection of the composite beam specimen was determined to be a full shear connection. The sectional and material characteristics of each composite beam specimen are summarized in Table 1. Figure 2 shows the test setup and instrumentation.

TABLE 1. LIST OF SPECIMENS WERE USED TO CARRY OUT THE CORROSION TEST

	Concrete slab			No. of transverse	Note
	Grad e	f <sub>c</sub>	Section (H×W)		
		MPa	mm		
B1	C30	39.9	120×600	2	All holes

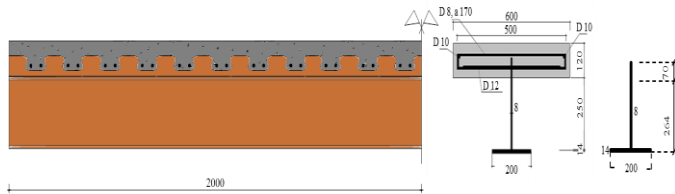


Fig. 1. Dimension and geometry of specimens SC-0

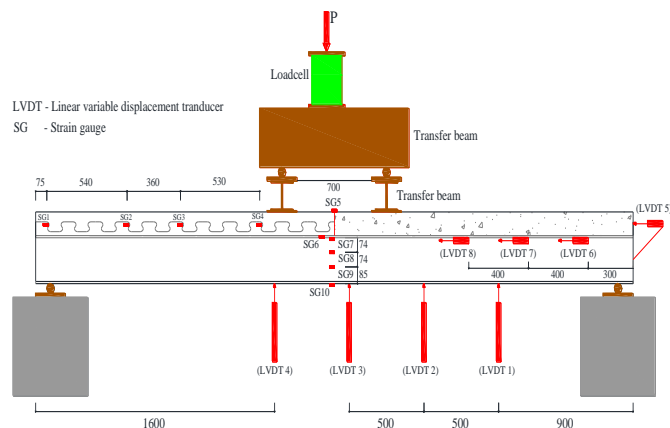


Fig. 2 Test schematic and arrangement of instruments

### B. Software, element types and mesh construction

The nonlinear finite element analysis software, with one of the most used nowa-days being ANSYS environment. The program offers a wide range of options regard-ing element types, material behavior and numerical solution controls, auto-meshers. The finite element types were chosen in the model of ANSYS 14. Accordingly, the finite element type of steel beam was (SHELL181), concrete slab was (SOLID65), shear connector was (COMBIN39), transfer plate and support were (SOLID45), and rebar was (LINK180 and REINF264). The characteristics of each type of element were shown in Fig. 3.

The von Mises yield criterion with isotropic hardening rule (multilinear work-hardening material) is used to represent the steel beam behavior, as shown in Fig. 10. The equation is given below

$$\sigma = f_y + E_h(\varepsilon - \varepsilon h) \left( 1 - E_h \frac{\varepsilon - \varepsilon h}{4(f_u - f_y)} \right)$$

where  $f_y$  and  $f_u$  are the yield and ultimate tensile stresses of the steel component, respectively;  $E_h$  and  $\varepsilon h$  are the strain hardening modulus and the strain at strain hardening of the steel component, respectively

The concrete slab behavior is modelled by a multilinear isotropic hardening relationship, using the von Mises yield criterion coupled with an isotropic work hardening assumption. The types of 3D failure surface of concrete according to Willam và Warnke. The concrete tensile strength and the Poisson's ratio are assumed as 1/10 of its compressive strength and 0.2, respectively. The concrete elastic modulus is evaluated according to Eurocode 4 (2004), is given below:

$$E_c = 9500(f_c + 8)^{1/3} (\frac{\gamma_c}{24})^{1/2}$$

Where  $\gamma_c$  is equal to 24 kN/m<sup>3</sup>.

In all analyses, the number/spacing of studs adopted in the experimental programmes is utilised. As far as the shear connector behavior is concerned, the load–slip curves for the newly puzzle shape of crestbond shear connectors are used (obtained from available push-out tests) by defining a table of force values and relative displacements (slip) as input data for the nonlinear springs. These springs are modelled at the steel–concrete interface.

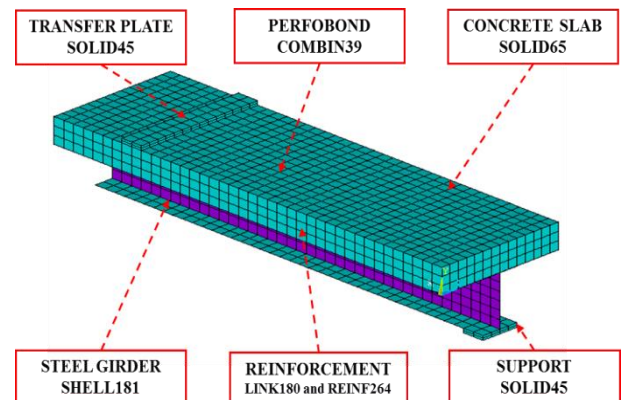


Fig. 3. Software, element types and mesh construction

### III. RESULTS AND DISCUSSION

### A. Comparison with experimental results

The results of four-point loading tests were compared to assess the behaviors of composite beams with the newly puzzle shaped of crestbond in relation to the number of transverse rebars and concrete strength. Table 4 summarizes

the loading test results of the composite beam specimens, including the deflections, relative slips at the composite interface, and strains of the concrete slab, steel beam, and crestbond rib. Simultaneously, the results of numerical analysis were also compared with ex-periment results.

#### B. Load-deflection relationships at the middle of composite beam

As can see that, the results from modeling and experiments are in agreement. The difference is approximate 0.83%, 0.33%, and 0.47% for B1, B2 and B3, respectively.

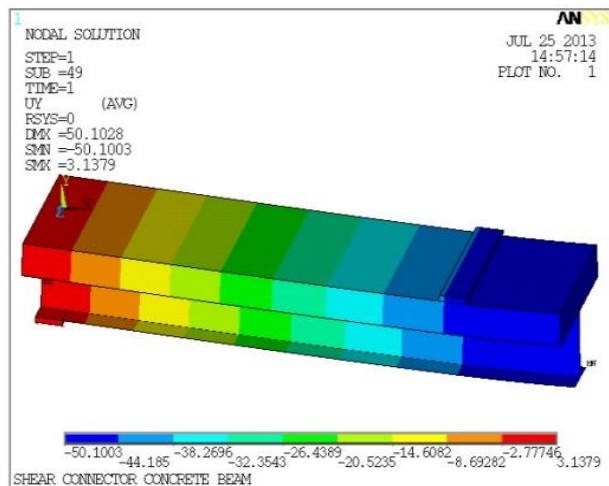
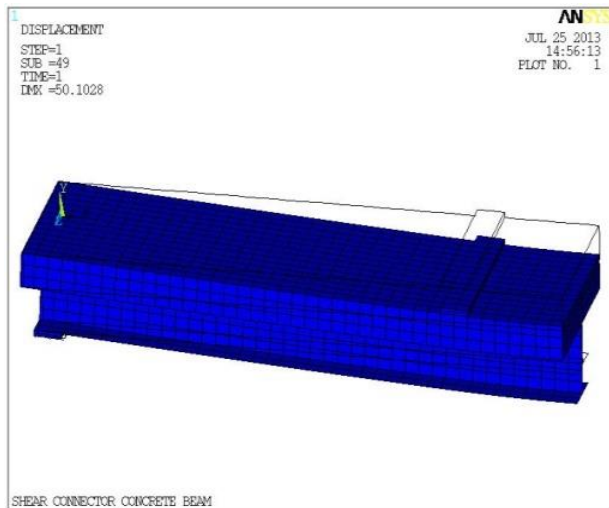
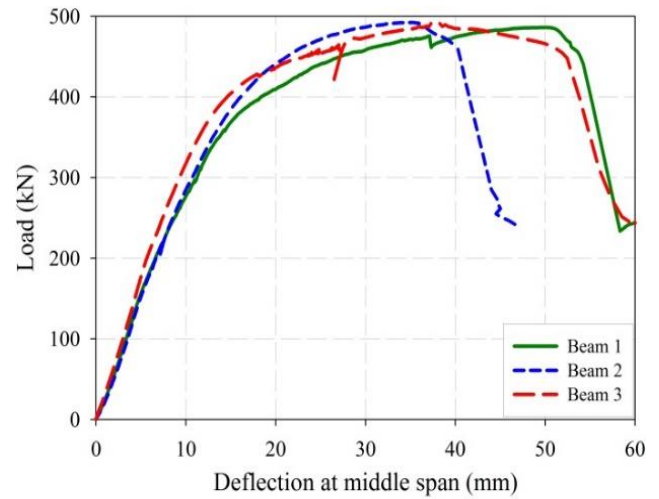
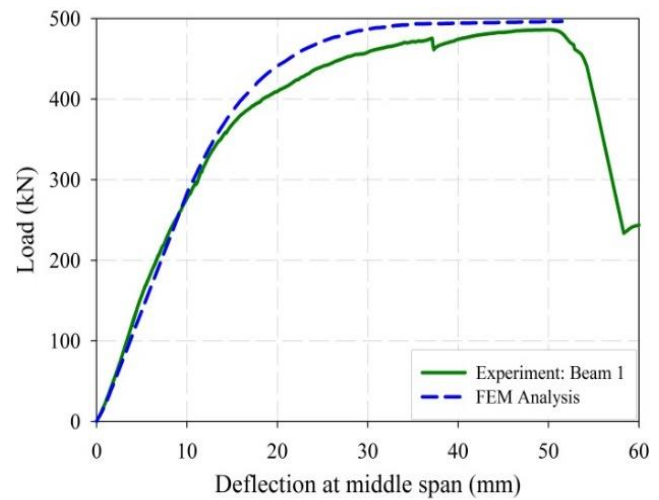


Fig. 5. Deflection of composite beam exported from Ansys



a) Experiment data



b) Beam 1

Fig. 6. Comparison of load-midspan deflection between experiment and FEM analysis

#### C. Failure mode

Figure 8 shows that only the neutral axis of beam B1 was located correctly, based on its designation.

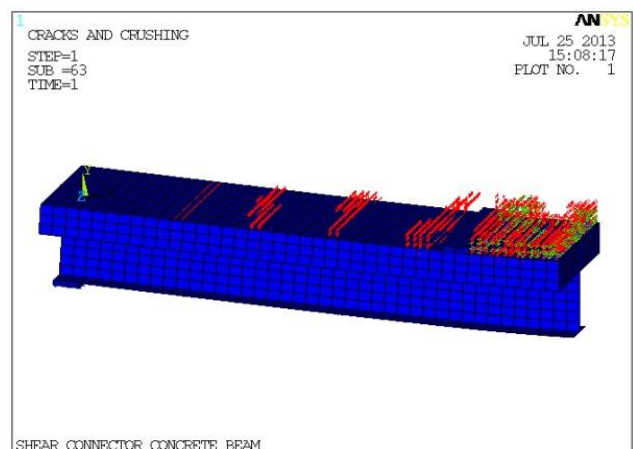


Fig. 7. Failure mode exported from Ansys



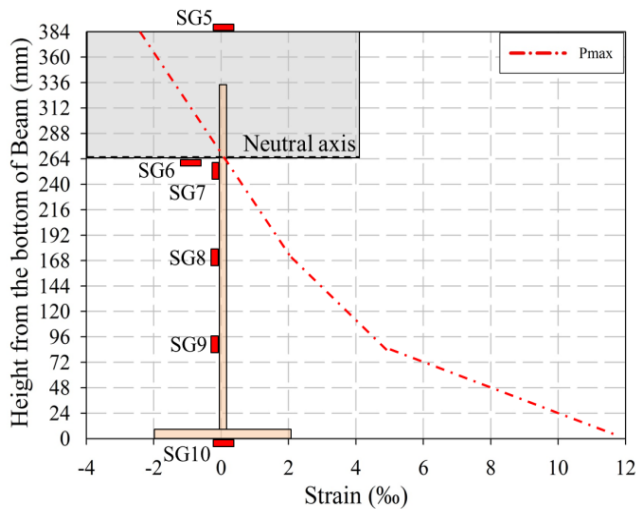


Fig. 8. Failure mode of composite beam due to FEM analysis

#### IV. CONCLUSIONS

A three-dimensional finite element model of composite beams is proposed based on the use of the commercial software ANSYS. All the main structural parameters and associated nonlinearities are included. The FEM analysis can be applied to re-place the experiment on the real composite beam based on the fully three-dimension FEM model. Comparisons with experimental results shown acceptable agreement between results obtained from modeling analysis and experimental push out tests. Effect of concrete compressive strength and concrete slab width are displayed in detail through deflection and relative slip. There is good in agreement between results of relative slip between concrete slab and steel girder obtained from modeling and

experimental push out test results. It is supposed that using element COMBIN39 is correct enough to reflect the nature of the newly puzzle shaped crestbond rib shear connector in composite structures.

#### REFERENCES

- [1] T. H. V. Chu, D. V. Bui, V. P. N. Le, I.T. Kim, J. H. Ahn, Duy Kien Dao, (2016), "Shear resistance behaviors of a newly puzzle shape of crestbond rib shear connector: An experimental study", *Steel and Composite Structures*, Vol. 21, No. 5, pp. 1157 - 1182 I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [2] V. P. N. Le, D. V. Bui, T. H. V. Chu, I.T. Kim, J. H. Ahn, Duy Kien Dao, (2016), "Behavior of steel and concrete composite beams with a newly puzzle shape of crestbond rib shear connector: An experimental study", *Structural Engineering and Mechanics*, Vol. 60, No. 6, pp. 1001 - 1019 R. Nicole, "Title of paper with only first word capitalized," *J. Name Stand. Abbrev.*, in press.
- [3] Duy Kien Dao, T. H. V. Chu, D. V. Bui, V. P. N. Le (2017). "Application of a newly puzzle shaped crestbond rib shear connector in composite beam using opposite T steel girder: An Experimental Study", *The 4th conference Congrès International de Géotechnique Ouvrages Structures CIGOS-2017M*. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.
- [4] T. H. V. Chu, V. P. N. Le, Duy Kien Dao, T. H. Nguyen, D. V. Bui, (2017). "Shear Resistance Behaviors of A Newly Puzzle Shape of Crestbond Rib Shear Connector: An Experimental Study", *The 4th conference Congrès International de Géotechnique Ouvrages Structures CIGOS-2017*
- [5] Nguyen, H. and Kim, S. (2009), "Finite element modeling of push-out tests for large stud shear connector", *Journal of Constructional Steel Research.*, 65(10-11), 1909-1920. <https://doi.org/10.1016/j.jcsr.2009.06.010>
- [6] Queiroz, F.D., Queiroz, G. and Nethercot, D.A. (2007), "Finite element modelling of composite beams with full and partial shear connection", *Journal of Constructional Steel Research.*, 63(4), 505-521. <https://doi.org/10.1016/j.jcsr.2006.06.003>

# A Mobile Deep Convolutional Neural Network Combined with Grad-CAM Visual Explanations for Real Time Tomato Quality Classification System

Loc-Phat Truong  
Faculty of Mechanical Engineering  
Ho Chi Minh city University of Technology  
and Education  
Ho Chi Minh city, Viet Nam  
15146227@student.hcmute.edu.vn

Bach-Duong Pham  
Faculty of High Quality Training  
Ho Chi Minh city University of Technology  
and Education  
Ho Chi Minh city, Viet Nam  
bachduong@hcmute.edu.vn

Quang-Huy Vu  
Faculty of High Quality Training  
Ho Chi Minh city University of Technology  
and Education  
Ho Chi Minh city, Viet Nam  
huyvq@hcmute.edu.vn

**Abstract**— This study develops a control system to classify tomatoes based on deep learning algorithms. In the main part, the CNN model was designed and trained to classify RGB image of tomato from 2D camera. In the hardware part, a conveyor system was design to implement and test the algorithms. This system contains a conveyor belt, pneumatic pistons, camera, embedded computer and its control circuit. In the third part, the CNN model was deployed into product through the embedded computer to interactive with actuators. To validate in practice, the system was tested to run in real-time and the authors measured the classify capability of this system. As the outcome, the system worked well with high speed and high accuracy, additionally, it is very intuitive with the visualization of model prediction.

**Keywords**— Computer vision, Deep learning, Convolutional neural network, Embedded system.

## I. INTRODUCTION

High quality agricultural product hugely contributes to the development of agriculture. People increasingly concentrate on improving the quality of their agricultural product, one particular example is the tomatoes. Many approach was developed to perform the classification of tomato quality. Very first approaches use color-based image processing in the RGB image [2], the newer is performed by unsupervised learning method for clustering [1], or with many features from multi-sensor fusion image [3], recently many researches start using machine vision to obtain more complex feature about the different of tomato and background [4], and latest is the approaches using the huge strength of data, deep learning algorithm [5,6].

Deep learning has been an outstanding technique in recent years. Convolutional neural network (CNN) is a type of neural network in deep learning applying mainly in computer vision. Since 2012, CNN has achieved many breakthroughs in machine vision, including image classification [8,9], object detection [10,11], or instance segmentation [12]. However, many current state-of-the-art CNN architectures for image classification have a huge computation cost to get very high accuracy leading to extremely low performance in real time process and be unusable in practice with mobile device or embedded computer. Additionally, even though CNN has unlocked extraordinary capability of computer, many researchers consider it as the black box, the empirical science because they have not yet had the decomposability into explainable and understandable components [7]. Therefore,

this study focused on building a CNN architecture which has very low computation resources based on Depthwise Convolutions and Bottleneck Residual Block [14] but still well fit with the dataset. In other side, the CNN model is also combined with the Gradient weights class activation mapping (Grad-CAM) for visual explanation powered by Ramprasaath et al [15] to be more intuitive by clearly answering the question “Where is it looking at?”.

The dataset of tomato is collected and processed with balance distribution introduced in Section II. Section III addresses intensively about the detail of model design process and system architecture. The hardware part is included a Raspberry Pi 3 model B plus quad-core CPU, a Logitech camera, a HDMI LCD screen, a STM32F1 microcontroller and an actuator system consisted conveyor, pneumatic cylinders, tomato container. The CNN model following by Grad-CAM [15] is brought in Raspberry Pi 3 using RGB image from Logitech camera placed on top of the conveyor. The classification result of CNN model is transferred to the STM32 microcontroller for controlling pneumatic cylinder, along with that, the attention map created by Grad-CAM algorithm is displayed on the HDMI screen. The compound work well and stably up to 7 frames per second (FPS) without any hardware optimization and software quantization in Raspberry Pi 3B plus is shown beside the training process result and evaluation in Section IV. The block diagram of system is described below in Figure 1. Eventually section V concluded the research and figured out the future work.

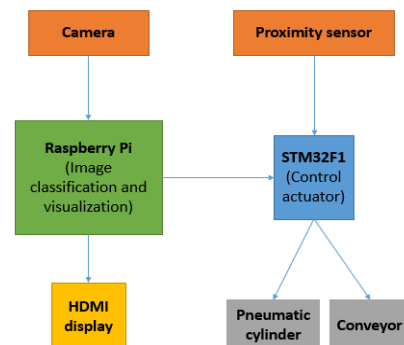


Fig. 1. Block diagram of system.

## II. DATASET

During this research, images of tomato have been captured into 6 categories which stand for 5 ripeness levels and a stage

of tomato non-appearance, nearly 6000 pieces, each category contains about 1000 images, this balance distribution makes the CNN model not biased the majority in dataset. In order to increase the diversity and generalization of dataset, images are taken in a various type of external condition, for instance the brightness or the position of tomato, or even adding noise object as paper, key, etc.

t-Distributed Stochastic Neighborhood Embedding (t-SNE) is an unsupervised learning algorithm in machine learning for visualization developed by Maaten et al [16]. It is a nonlinear dimensionality reduction technique popular used to encode high-dimensional data into a low-dimensional data of 2 or 3 dimensions, which can be easily plotted on chart for intuition about data. Figure 2 is a result of a part of dataset processed by t-SNE, we can see the relation of all pieces in dataset after visualizing.

### III. SYSTEM ARCHITECTURE

#### A. Hardware platform

- Raspberry Pi 3B Plus

Raspberry Pi 3B Plus is a single-board computer with full functions as a normal desktop. This computer is equipped a faster 64-bit 1.4 GHz quad core processor ARM Cortex-A53 CPU (BCM2837) and 1 GB LPDDR2 SDRAM. Raspberry Pi is operated by Raspian OS which is in Linux family, thus, it is easily using for developers who are familiar to Unix-like OS, additionally, many frameworks for deep learning have been supported on Raspian currently, this is a strong plus for using this computer to develop deep learning applications.

- Camera Logitech C270

Logitech is a popular brand in camera production. Logitech C270 model is a very cheap camera with capability of 720p video streaming through a high speed 2.0 USB port and give us 640×480 frames at 30 FPS.

- Microcontroller STM32F103C8T6

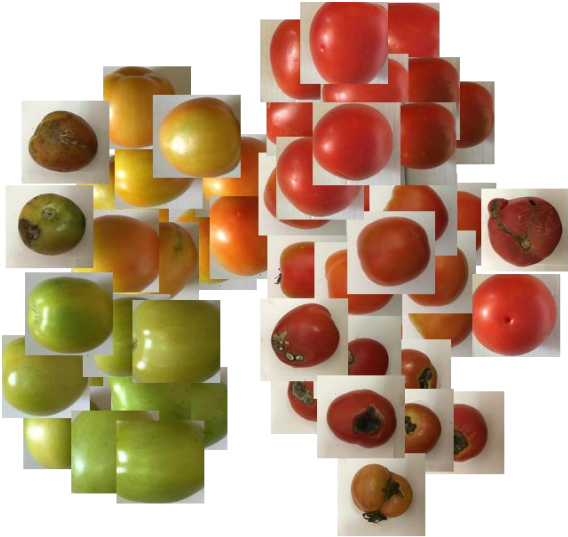


Fig. 2. t-SNE visualization of a part of the tomato dataset

STM32F103C8T6 is a high speed 32-bit microcontroller with clock boost up to 72 MHz. There are 37 I/O pins on this chip, a large program memory size with 64 KB and many powerful peripherals like PWM, DMA, etc. It is used as the

center controller due to the capability of high processing, flexible connectivity with actuator and high speed communication protocol.

#### B. CNN architecture

- Depthwise Convolution

Depthwise convolution is a new type of computation which is lower computation cost from standard convolution. It is used as a key component in many network architectures [13,17].

Standard convolution takes a feature map in the shape of  $W \times H \times D$  as input to apply with  $D_r$  kernels  $K \in R^{k \times k \times D}$  and the output feature map is produced in the shape of  $W_r \times H_r \times D_r$ . It takes the computation cost of  $W_r \cdot H_r \cdot D_r \cdot D \cdot k \cdot k$ .

Depthwise convolution just uses  $D$  kernels  $K \in R^{k \times k \times 1}$  to produce the output feature map in the shape of  $W_r \times H_r \times D$  and usually is followed by an additional pointwise convolution which actually is a standard convolution with  $D_r$  kernels  $K \in R^{1 \times 1 \times D}$  to perform the output feature in the shape of  $W_r \times H_r \times D_r$ . This combination just takes computation cost of  $k \cdot k \cdot W_r \cdot H_r \cdot D + D_r \cdot D \cdot W_r \cdot H_r$ , it is called Depthwise separable convolution that is invented to replace standard convolution for acceleration because of the same output feature map but less computation cost. However, in this research we just only use depthwise convolution but not depthwise separable convolution to make the bottleneck residual block explained clearly in the next section.

- Bottleneck residual block

As mentioned in MobileNet V2 [14], bottleneck is able to extract a reasonable amount of information for many tasks while it is designed to reach the memory efficiency. Bottleneck residual block includes 3 component layers:

The first layer is a standard convolution layer which applies  $k$  kernels  $1 \times 1$  followed with ReLU6 activation function to transform nonlinearly the input feature map  $F \in R^{w \times h \times d}$  into a new feature map  $F_1 \in R^{w \times h \times k}$ , where  $k = t \cdot d$ ,  $t$  is called the expansion factor. Therefore, this layer is called the expansion.

The second layer is a depthwise convolution layer that uses optional kernel size followed again with ReLU6 activation function, the  $F_1$  is transformed nonlinearly each channel into the feature map  $F_2 \in R^{w \times h \times k}$ .

The last layer is a standard convolution layer using  $d$  kernels  $1 \times 1$  to compress in a linear transformation the  $F_2 \in R^{w \times h \times k}$  into the  $F_3 \in R^{w \times h \times d}$ . The  $F_3$  has the same shape with the  $F$ , thus we add a shortcut between  $F$  and  $F_3$  due to the intuition of classical residual connections [18] that the capability to impact to the gradient through many layers in network.

- Network architecture

In this section, we will discuss more about the detail of model architecture. Our CNN architecture is designed primarily based on the bottleneck residual block because of the target of optimizing inference speed without hurting the performance of model. The detail of our bottleneck residual block is shown in Figure 3. All bottleneck residual blocks used in our architecture have the expansion factor  $t$  of 2, we found that the larger factor will contribute a higher

performance but hurt too much the inference speed that it cannot be used in an embedded computer in practice.

Our CNN model shown in Table 1 is initialized with a standard convolution layer with 32 kernels  $1 \times 1$  and a max pooling layer to reduce the half size of feature map, followed by 2 bottleneck residual block described above, and we specify the batch normalization between each layer plus the average pooling layer and the global average pooling before going to 2 fully connected layer at the end. The motivation of using average pooling layer instead of max pooling is because there are too few parameters in this type of model architecture due to the purpose of inference speed, the authors see that it is not reasonable if we eliminate a big amount of weights by max pooling (just keep the maximum value of each window), therefore we use average pooling instead to let all weights affect to the model. At total, our CNN model has 10694 trainable parameters and occupies 252 Kb on disk.

- Training strategy

Logitech Because our model is a categories classification model, we apply the cross entropy loss [20] for many categories to our training process

$$J = \frac{-1}{N} \sum_{i=0}^N \sum_{k=0}^M y_{i,k} \cdot \log(\hat{y}_{i,k}) + (1 - y_{i,k}) \cdot \log(1 - \hat{y}_{i,k}) \quad (1)$$

Where N is the number of samples and M is the number of categories,  $\hat{y}$  is stand for the model prediction and  $y$  is the ground truth label.

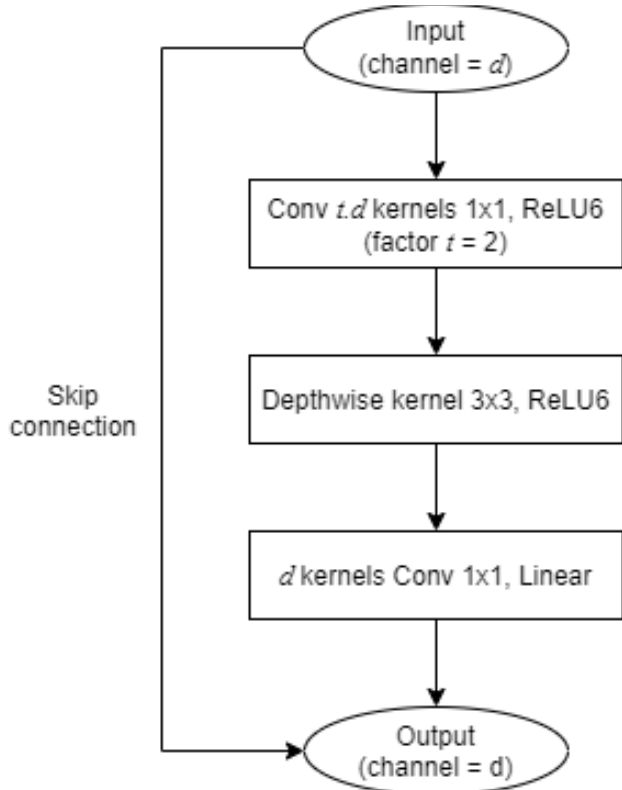


Fig. 3. Bottleneck residual block architecture used in our network

TABLE. I. NETWORK ARCHITECTURE

Input Size	Operation	Detail
$100 \times 100 \times 3$	Conv2D $1 \times 1$	k = 32 strides = 1 padding = same
$100 \times 100 \times 32$	BN	default
$100 \times 100 \times 32$	Max Pooling	kernel $2 \times 2$ strides = 2
$50 \times 50 \times 32$	BRB	$t = 2$
$50 \times 50 \times 32$	Average Pooling	kernel $2 \times 2$ strides = 2
$25 \times 25 \times 32$	BRB	$t = 2$
$25 \times 25 \times 32$	GAP	default
32	Dense	ReLU6
6	Softmax	default

Note: The architecture use primarily the bottleneck residual block (BRB) with  $t = 2$ , additionally the authors also apply batch normalization (BN) for regularization and weights stability, average pooling layers and global pooling layers.

The Adam optimizer [21] is manipulated to update weights after each training iteration, the detail is expressed in (2), (3), (4).

$$v_t = \beta_1 \cdot v_{t-1} - (1 - \beta_1) \cdot \frac{\partial}{\partial \omega_t} J(\omega_t) \quad (2)$$

$$s_t = \beta_2 \cdot s_{t-1} - (1 - \beta_2) \cdot \frac{\partial}{\partial \omega_t} J(\omega_t)^2 \quad (3)$$

$$\omega_{t+1} = \omega_t - \eta \frac{v_t}{\sqrt{s_t + \varepsilon}} \cdot g_t \quad (4)$$

Where  $\eta$  is the initial learning rate,  $\beta_1$ ,  $\beta_2$ ,  $\varepsilon$  are the hyperparameters.

### C. Gradient-weights Class Activation Mapping (Grad-CAM) Visualization

- Grad-CAM

To comprehend why neural network make the prediction, the authors apply a technique called Gradient-weighted Class Activation Mapping (or Grad-CAM) which is published in 2017. According to [15], this technique produces “visual explanations” for large CNN-based model, make them more transparent.

It uses the gradients of any target concept (say logit is the model prediction of tomato kinds in our case), following into the last convolution layer to produce a localization map highlighting the important regions in an image that impact hugely to model prediction.

In order to obtain the class discriminative localization map Grad-CAM  $L_{Grad-CAM}^c \in \mathbb{R}^{u \times v}$  of width u and height v for any class c, we compute the gradient of the class score  $y^c$  with respect to feature maps  $A^k$  of convolutional layer, i.e.  $\frac{\partial y^c}{\partial A^k}$ . Then we global average pooled these gradients to obtain the neuron important weights  $\alpha_k^c$ :

$$\alpha_k^c = \frac{1}{Z} \sum_{i=0}^u \sum_{j=0}^v \frac{\partial y^c}{\partial A_{ij}^k} \quad (5)$$

Eventually,  $L_{Grad-CAM}^c$  is obtained by weights  $\alpha_k^c$  linear combining with activation map  $A^k$  and a following ReLU activation:



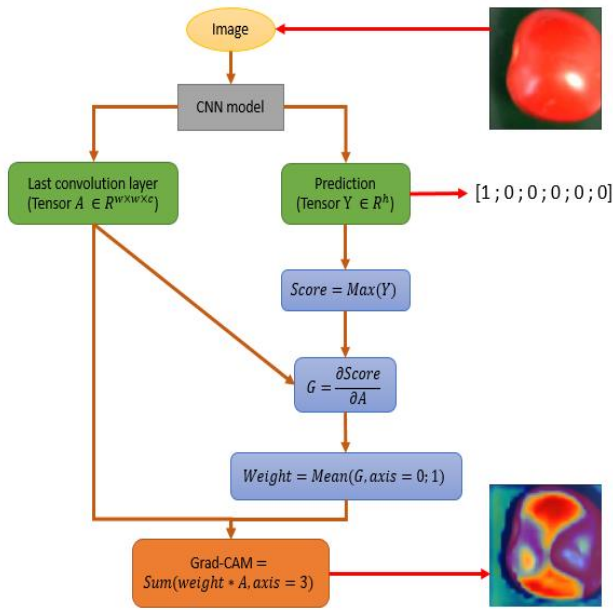


Fig. 4. Flowchart of combination algorithm for each frame in real time process loop

$$L_{Grad-CAM}^C = ReLU \left( \sum_k \alpha_k^C \cdot A^k \right) \quad (6)$$

- CNN combined with Grad-CAM in real time process

For a deeper and more comprehensive intuition about the model, additionally for more general evaluation and better visualization of the model prediction, we stacked our model with the Grad-CAM behind to get a heat map that can represent the model attention on every region in image to make its prediction. The pipeline is shown in Figure 4, every frame got from camera in real time is put into this pipeline to make the prediction and visualize the attention map.

#### IV. EXPERIMENTS

##### A. Training Process

In the section II, we discussed about the approach for collecting data and the structure of dataset, 6 categories with nearly 1000 instances per category. Totally, our dataset has 6441 images and it is divided into 3 subsets [19] which are necessary for training machine learning model and modifying its hyperparameters: train set - 80%, test set - 10% and cross validation set - 10%. The images of each category is gathered into 1 folder named similar with the name of this category and the label is simply assigned to image using name of the folder by an automate Python script.

We used Google Colaboratory which is a free and accelerating environment for training process. It is powered by the GPU Tesla T4, Ram 2 Gb, although Google Colab was warned that it usually interrupts incidentally because of the free price but it is still enough for training a small model within below 1 day. Due to the purpose of running classifier in real time on a Raspberry, the images in dataset is resized to the shape of  $100 \times 100$  as well as the input shape of CNN model.

We trained CNN model with batch size of 32 and initial learning rate of 0.001, the hyperparameters of Adam optimizer was set default with  $\beta_1$  of 0.9,  $\beta_2$  of 0.999, no learning rate schedules and common  $\epsilon$  of  $10^{-7}$ . The model was trained by

300 epochs which take us about 2 hours, because the model is very small due to the design purpose, so it took less time to train than the other current state of the art for tougher task. Eventually, we got the classification loss of below 0.01 on both train set and validation set with no overfitting problem shown in the Figure 5. The model totally converges after step 250<sup>th</sup>.

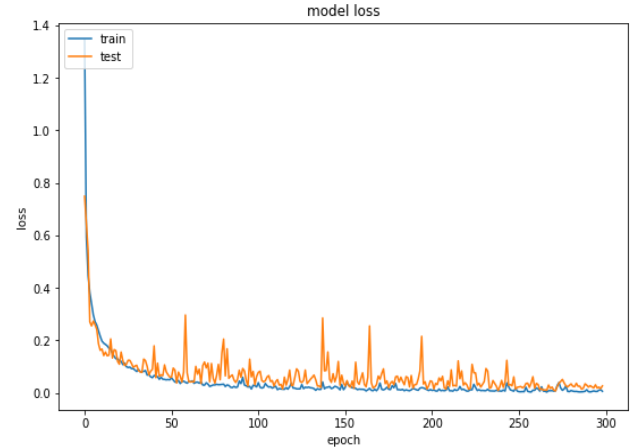


Fig. 5. Training and Validation loss graph

##### B. Real Time Evaluation

Besides monitoring the loss value and accuracy of model during training, it is necessary to evaluate the model behavior with test set and especially in practical environment. After training on Google Colab, the model is embedded on Raspberry Pi to test the ability of working in reality. Combined with Grad-CAM visualization, the pipeline is fully evaluated in real time. Tomato is put into conveyor mechanical model shown in Figure 6. The original CNN model has the speed of 0.067 s/frame on Raspberry Pi 3B+, however after the combination with Grad-CAM step, the speed of model currently is reduced to a half, 0.143 s/frame, but 0.143 s/frame (can be said as 7 FPS) still is an acceptable number for real-time processing. Some test result is shown in Table 2, the first column contains the input frames from camera, the next column is the prediction and the confident score of model, the last column represents the corresponding attention map of model with the input image.

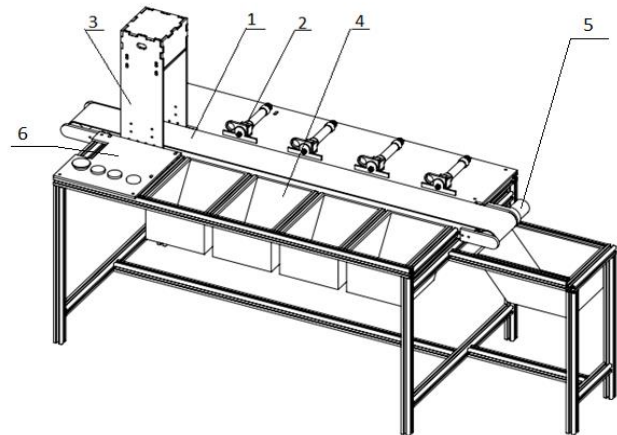

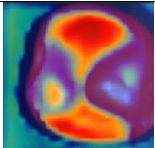

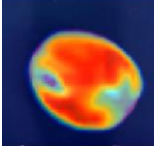

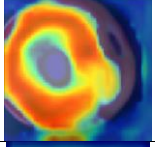

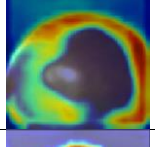

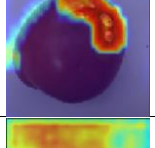

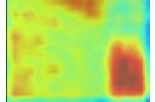


Fig. 6. Mechanical model with 6 main parts:  
1 – conveyor, 2 – pneumatic cylinder, 3 – camera box,  
4 – tomato container, 5 – motor, 6 – LCD screen.



TABLE. II. RESULT IN PRACTICAL ENVIRONMENT

Input	Output	
	Result	Attention Map
	Red ripeness: 0.984102	
	Orange ripeness: 0.999183	
	Yellow ripeness: 0.990013	
	Green ripeness: 0.945731	
	Corrupted: 0.992356	
	Non tomato: 1.000000	

## V. CONCLUSION

In this research, we clustered tomato into 6 categories including 5 ripeness period and 1 stage of the non-appearance tomato. Depending on that, we designed a super light-weight convolutional neural network mostly motivated from MobileNet V2 to perform the classification task with high inference speed and acceptable accuracy, which totally takes 252Kb of hard disk. Sequentially, we manipulated the Grad-CAM algorithm to visualize our network prediction into a heat map which help us have a deep insight about our CNN model.

Before training model, we totally built the dataset by hand, collected image and preprocessed them. In order to monitor the distribution of dataset, we applied the t-SNE on our dataset, and we were able to delete bad images which can make noise to the model. One important issue which we need to pay attention to is the overfitting problem, thus we divided our dataset into 3 subsets: train, test and validation. In order to avoid overfitting, the loss on validation set is the key point which need to monitor during training to modify the hyperparameters.

Eventually, the CNN model after training which have combined with Grad-CAM algorithm was deployed on a small embedded platform Raspberry Pi 3 and reached 7 FPS at total.

## REFERENCES

- [1] H. Yin, Y. Chai, S. X. Yang and G. S. Mittal, "Ripe Tomato Recognition and Localization for a Tomato Harvesting Robotic System," *2009 International Conference of Soft Computing and Pattern Recognition*, Malacca, 2009, pp. 557-562.
- [2] Y. Gejima, Houguo Zhang and M. Nagata, "Judgment on level of maturity for tomato quality using L\*a\*b\* color image processing," *Proceedings 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003)*, Kobe, Japan, 2003, pp. 1355-1359 vol.2.
- [3] M. Li and H. Wu, "Target Identification and Growth State Distinguish of Tomato," *2010 World Automation Congress*, Kobe, 2010, pp. 357-364.
- [4] L. Xiao-lian, L. Xiao-rong and L. Bing-fu, "Identification and Location of Picking Tomatoes Based on Machine Vision," *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, Shenzhen, Guangdong, 2011, pp. 101-107.
- [5] E. Suryawati, R. Sustika, R. S. Yuwana, A. Subekti and H. F. Pardede, "Deep Structured Convolutional Neural Network for Tomato Diseases Detection," *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Yogyakarta, 2018, pp. 385-390.
- [6] L. Zhang, J. Jia, G. Gui, X. Hao, W. Gao and M. Wang, "Deep Learning Based Improved Classification System for Designing Tomato Harvesting Robot," in *IEEE Access*, vol. 6, pp. 67940-67950, 2018.
- [7] Lipton, Zachary C. "The mythos of model interpretability." *Queue* 16.3 (2018): 31-57.
- [8] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." *Thirty-first AAAI conference on artificial intelligence*. 2017.
- [9] Tan, Mingxing, and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *arXiv preprint arXiv:1905.11946* (2019).
- [10] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).
- [11] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.
- [12] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [13] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [14] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [15] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [16] Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), pp.2579-2605.
- [17] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [19] Reitermanova, Z. "Data splitting." *WDS*. Vol. 10. 2010.
- [20] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19-67, 2005.
- [21] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

# Design of Driver Circuit to Control Induction Motor Applied in Electric Motorcycles

Dinh Cao Tri

Faculty of Vehicle and Energy Engineering  
HCMC University of Technology and Education  
No.1 Vo Van Ngan Str., Linh Chieu Ward, Thu Duc District,  
Ho Chi Minh City, Vietnam  
dinhcaotri1302@gmail.com

Le Thanh Phuc

Faculty of High Quality Training  
HCMC University of Technology and Education  
No.1 Vo Van Ngan Str., Linh Chieu Ward, Thu Duc District,  
Ho Chi Minh City, Vietnam  
phuctl@hcmute.edu.vn

**Abstract** — This paper presents a design of a controller circuit and a control method for a three-phase asynchronous motor applied in electric motorcycles. The research focused on two main tasks including signal processing and power driving. In particular, the signal control device is an Atmega328P microcontroller, which is basically an Advanced Virtual RISC (AVR). The microcontroller plays a role as a central processor to create the Pulse Width Modulation (PWM) and turn the DC voltage on and off. The width of each pulse was altered based on AVR so that the overall voltage at the output is similar to the sine wave. Insulated-gate bipolar transistors (IGBTs) were used as the power part to drive the induction motor. The control method is based on the principle of variable frequency drive (VFD). The DC voltage is converted to an AC sine wave voltage by reading the lookup table. The waveform can be refined until closely resembles that of a pure sine wave. The experiment then was performed to evaluate the circuit. The driver is applied to control a 2HP motor in an electric motorcycle model.

**Keywords**— inverter, sinusoidal wave, PWM, electric motorcycle.

## I. INTRODUCTION

In the face of increasing environmental pollution, many countries around the world encourage people to use eco-friendly vehicles [1][2] and Vietnam is certainly not out of the global trend. Currently, Vietnam's air pollution exceeds the permissible level that can adversely affect human health, which is mainly caused by emissions from vehicles [3]. In fact, Vietnam is a nation that depends heavily on motorcycles because roads in Vietnam have many nooks and crannies that are not convenient for driving cars, but motorcycles can do this thanks to its flexibility, so motorcycles are always in high demand in the near future. In other words, replacing motorcycles using gasoline fuel to electric motorcycles is the most feasible. Riding electric motorcycles is an important solution not only to show civilization, but also to see it as a paramount thing to protect your own living environment and the community as well.

However, electric motorcycles still have not captured the taste of most consumers in Vietnam indeed. There are many reasons to explain this and one of them is the cost problem. Because now, in the stage of globalization, electric vehicle manufacturers are shifting production investment abroad in order to maximize material resources and manpower supply from all over the world. Nevertheless, this is the main reason that directly affects shipping costs and product prices due to geographical distance.

There are many studies related to the PWM technique of induction motor speed control which is based on the VFD principle. For instance, Jamadar, Kumbhar, Gavane, Suttrave [4] presented the circuit of an induction motor using a PIC microcontroller. Kojabadi [5] analyzed and compared different PWM methods to minimize distortion and harmonic noise. Liang, Laughy, Liu [6] calculated the parameters for motor starting at low frequency from 2-10Hz based on cross-line and VFD methods. Most of the researches mainly analyzed the control techniques, optimized voltage and current for each operating mode of the motor so as to improve efficiency. However, these papers did not mention the calculation and design of a driver circuit in detail.

In this context, the induction motor will be chosen to replace the internal combustion engine because it is completely manufactured and assembled in Vietnam. Hence, the cost of products will be reduced, customers can easily access repair and maintenance services. In addition to motors, electric motor drive circuits play an indispensable part in the motor operation. Specifically, the driver circuit has the main task of capturing input signals (throttle position, vehicle speed) for calculation and processing, then transmitting the output signal to the actuator. Such settings input parameters must be computed based on characteristic curves of each of motor types in order to ensure some issues such as changing traction in accordance with the multiple different loads on the road surface or maximize acceleration. Furthermore, a driver circuit also requires choosing and arranging electronic components appropriately to reduce operation noise and increase circuit stability.

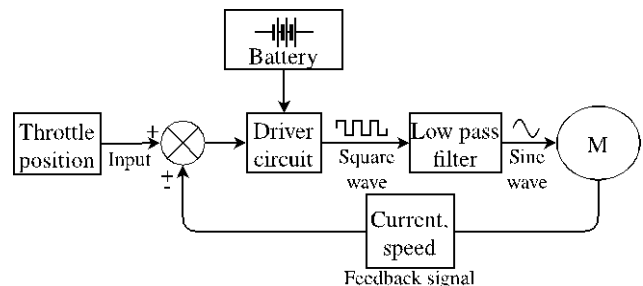


Fig. 1. The schematic of general system

The layout of this paper is written as follows. Section II represents an overview of the control method. Section III is about utilizing a microcontroller to create sinusoidal PWM (SPWM). Section IV mentions the design of a driver circuit. The experiment setup and results are in section V and section VI respectively. Lastly, section VII will summarize this paper.

## II. THE OVERVIEW OF THE CONTORL METHOD

### A. Voltage per frequency (V/f) control mode for induction motors

The actual speed of an electric motor depends on the operating frequency, the number of poles, and the relative slip between the stator and the rotor, is shown in (1) [7]. In which, the number of poles is a hardware that cannot be changed during the operation, the motor speed will now depend on the slip and the operating frequency. In other words, the actual speed of the electric motorcycle can be changed based on the change of frequency that is applied in the motor with the expectation that the motor will reach the desired speed corresponding to the lowest slip.

$$N = \frac{120f}{p} (1 - s) \quad (1)$$

However, changing the frequency while maintaining the constant amplitude of motor voltage will cause an increasing of current supplied to the motor. Basically, the current flowing into the motor has two main tasks, one is used to generate magnetic flux, called magnetizing current while the order is used to cause loss, reduce the lifetime of the motor. On the other hand, when the current is raised to a certain level, the magnetic saturation phenomenon will occur, so the magnetizing current cannot increase anymore. At this time, if the supply current continues to increase, it has the sole task of generating losses, lead to overheating. Conversely, if the flux falls below the norm that will reduce the motor torque. Thus, reducing the supply frequency to the motor is below the fundamental frequency often associated with a decreasing of the supply voltage to the motor that follows the constant ratio  $V/f$ . In other words, changing this ratio will keep the motor operating in a linear region with constant flux.

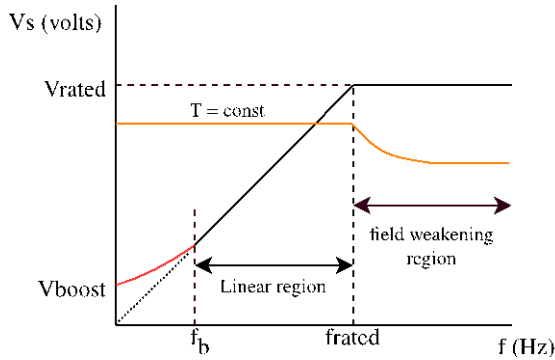


Fig. 2. Voltage and frequency variation

However, at low frequencies ( $f < f_b$ ), the effect of stator resistance in reducing the flux becomes very clear [8]. So, the minimum frequency is much less than the pullout torque at higher frequencies, and this could be a problem for loads that require a high starting torque at the low frequency band [9]. The low-frequency performance can be improved by raising the  $V/f$  ratio at low frequencies in order to keep full flux, a technique which is referred to a “low-speed voltage boosting” (the red line in figure 2) [10].

On the other hand, when the motor reaches a high speed that equals to the rated, the frequency  $f$  can be encouraged to push away from the  $f_{rated}$  ( $f > f_{rated}$ ) while maintaining the rated voltage to increase the motor speed. At this time, the motor operating in the area has a reduced magnetic flux but still ensures stability if the load is not changed abnormally.

### B. Sinusoidal PWM analysis

To generate an AC voltage using the SPWM method, a high-frequency triangle pulse signal was compared with a standard sine wave of frequency. If this control pulse is supplied to a single-phase inverter, the output will receive a PWM voltage of a frequency that equals to the sample sinusoidal frequency. The first harmonic amplitude depends on the power source supply and the ratio between the sample sine wave amplitude and the carrier wave. If the modulating sine wave voltage is greater than the carrier voltage, the PWM signal at the output will show a high level. Conversely, if the modulating sine wave voltage is lower than the carrier voltage, the PWM signal will be low. Figure 3 below shows the comparison between the carrier wave (orange line) and the modulated wave (red line), the result of this comparison is to create the high and level at the output signal. In particular, the comparison of a signal between the sinusoidal wave and triangle wave will be made by a microcontroller. The microcontroller main task is to generate sine and triangular waves, then combine them. The output is a PWM waveform.

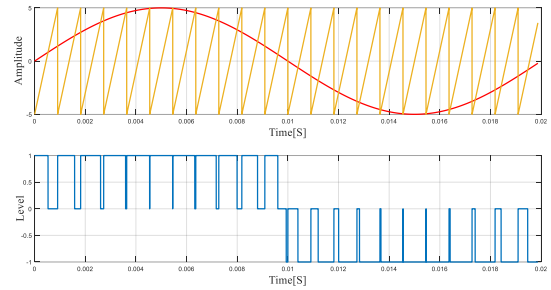


Fig. 3. Signal comparison between the carrier wave and modulated wave

## III. UTILIZING MICROCONTROLLER TO CREATE SPWM

### A. Initial carrier wave

The carrier frequency must be higher than the frequency of the modulating sine waves in order to ensure the output signal brings to the load must be as smooth as possible. Thus, the carrier frequency should be normally around 4-15KHz, which depends on the different load modes and the switching speed of a three-terminal power semiconductor device. In particular, if the carrier frequency is too high, it will create a hissing noise in the engine and making semiconductor devices become overheating. By contrast, if the carrier frequency is too low that will lead to the output signal is unstable.

The carrier frequency ( $f_{carrier}$ ) is set in the Atmega328P microcontroller by initializing the value for the TCNT0 register, which is the 8-bit register of timer 0. There are 3 types of PWM mode in Atmega328P including Clear Time On Compare Match (CTC) mode, Fast PWM mode and Phase Correct PWM mode, respectively [11]. To simplify the setting of the carrier frequency, if the wire connected from the motor to the inverter is short and the resistance on the wire is approximately zero, then the carrier frequency will be able to set to an average of 4 -15KHz.

$$f_{carrier} = \frac{f_{osc}}{N(2^8 - 1)} \quad (2)$$

Where  $f_{osc}$  is the frequency of the crystal that is set into the microcontroller. Commonly,  $f_{osc}$  should be set at the

highest 16MHz. Moreover,  $N$  variable represents the pre-scale factor (1, 8, 64, 256 or 1024) [11]. In addition, it is possible to change the carrier frequency during motor operation in the case of compulsory by increasing or decreasing the initial value of register TCNT0 (TCNT0 = 0 in standard).

Figure 4 [11] below shows how to generate the PWM. Register TCNT0 has the function of counting from 0x00 to 0xFF, the value of register TCNT0 will increase by one unit after each clock cycle. On the other hand, the value of OCRnx is used to compare with register TCNT0. For instance, in the “non-inverter” mode, when the TCNT0 value is equal to the OCRnx value, the output voltage will drop to low, then after the end of one cycle, the output voltage will be set to high. Repeating the process of changing the voltage level together with the variation of OCRnx value will generate square pulses of the desired frequency and width

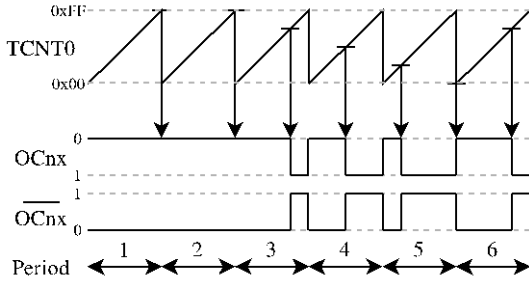


Fig. 4. Describe the function of the TCNT0 register

### B. Sinusoidal lookup table

The modulating wave is generated by creating a lookup table of values from 0 to  $\pi$  representing half of the period of the standard sine wave. As presented above, a modulating wave carries information about the root-mean-square (RMS) voltage and the frequency of the output ( $f_{output}$ ). Therefore, the change in amplitude and frequency of the modulated wave will lead to the change in amplitude and the frequency of the output voltage changes.

Firstly, the frequency of the sinusoidal modulated wave varies depending on the number of elements ( $N_{elements}$ ) in the lookup table. Furthermore, the range of the elements in the lookup table is from 0 to 255 correspond to the OCRnx value used to compare with the carrier. OCRnx will be updated to the next value of the element in the table after each loop. After updating all the element values in the table (semi-cycle from 0 -  $\pi$ ), this OCRnx register will switch to inverter mode (figure 5) to continue performing the next half-cycle ( $\pi - 2\pi$ ). The higher the modulation wave frequency, the shorter the time it takes to complete the cycle, then the less the number of elements in the lookup table and vice versa. The relationship between the number of elements in the lookup table and the frequency of the sinusoidal modulated wave is determined in (3). Note that a number of elements in the lookup table has to be a positive integer and divisible by 3 because the induction motor has three phases with each phase deviation by 120 degrees. Additionally, the elements in the lookup table should be distributed according to the sine wave formula to ensure that the modulated wave is close to the standard sine wave

$$f_{output} = f_{carrier} \frac{1}{2N_{elements}} \quad (3)$$

Additionally, the elements in the lookup table should be distributed according to the sine wave formula to ensure that the modulated wave is close to the standard sine wave.

$$OCRnx_i = 255 \sum_{i=0}^{N_{elements}} \sin\left(\frac{\pi}{N_{elements}} i\right) \quad (4)$$

Secondly, at each given frequency always get a corresponding voltage value according to the  $V/f$  ratio. Moreover, the output voltage is proportional to the amplitude of the modulated wave. For a DC voltage, the average voltage ( $V_{AV}$ ), the peak voltage ( $V_P$ ), and RMS voltage can be seen as one. But in a sine voltage, RMS voltage represents the standard voltage, which is shown in figure 5.

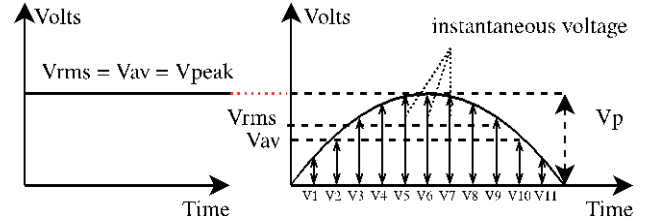


Fig. 5. Different voltage levels of DC voltage on the left side and sine voltage on the right side

The RMS voltage is determined by taking the square root of the sum of the squared instantaneous voltage values all over the total number of samples taken in each cycle [12]. Moreover, the number of samples corresponds to each of  $N$  elements of the lookup table. In which, the average voltage of each PWM pulse is the instantaneous voltage at a time. Therefore, the output sine wave voltage can be computed based on the value of each element of the table.

$$V_{out} = \frac{V_{peak}}{255} \sqrt{\frac{1}{N_{elements}} \sum_{i=0}^{N_{elements}} OCRnx_i^2} \quad (5)$$

### IV. THE DESIGN OF DRIVER CIRCUIT

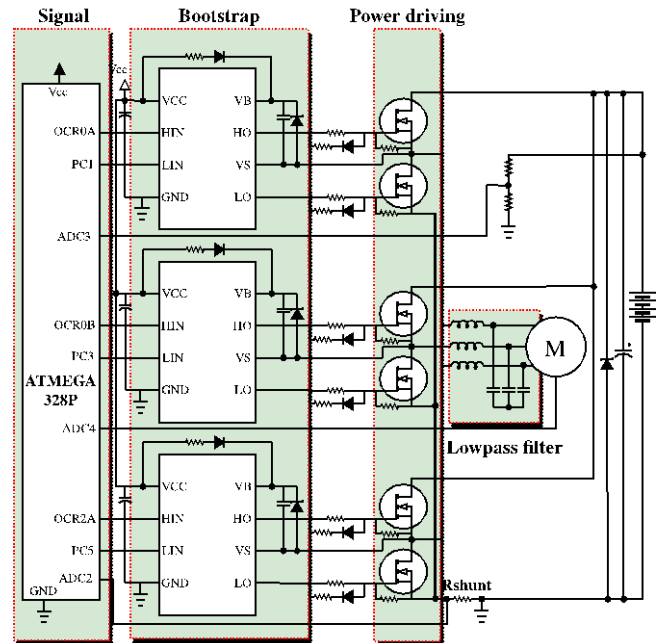


Fig. 6. General schematic of a driver circuit



A complete three-phase inverter circuit is fully displayed in the circuit diagram above. The circuit has three main parts: the control unit, the bootstrap unit and the power device element. Depending on the power of the motor or the external factors that the circuit can be modified to the accord.

### A. A bootstrap circuit and basic calculation

The main task of the bootstrap circuit is to control the power elements at the high side and low side thanks to three basic components: the bootstrap capacitor, the bootstrap diode and the resistor [13]. In particular, when the high side IGBT is off the low-side IGBT is on, the VS pin is immediately pull down to the ground by flowing through the collector to emitter of the power device. Then, the current from voltage supply charges the capacitor bootstrap through diode bootstrap (as the dashed-line path in figure 7 below). Next, when the low-side IGBT is off, the energy in the bootstrap capacitor will release to the high side IGBT from VB pin to HO pin, then the high side IGBT will be on.

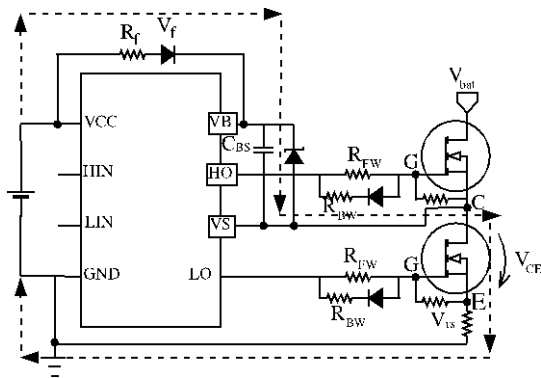


Fig. 7. The bootstrap circuit for a single-phase

The capacitor is calculated by the quotient of its  $Q$  charge and the  $U$  voltage applied to it. Besides, the quantity of charge that flows through a surface is the current intensity flowing through that surface per unit time. So, the total charge of the bootstrap capacitor is the total of the electric component charges passing through it, and the voltage that is applied to the capacitor is the  $V_{bs}$  voltage.

$$C \geq \frac{Q_{total}}{\Delta V_{bs}} \quad (6)$$

Where  $\Delta V_{bs}$  is a minimum drop voltage that can turn power device IGBT on at the high side, it can be determined by taking the voltage from the supply minus the component voltages involved in the capacitor charging process. as can be seen the path charging of bootstrap capacitor in figure 7, The current from the power supply flows through the bootstrap diode, the bootstrap capacitor, then flows through the CE terminal of the low side power device and finally passes through the resistor to recognize the current and then to the ground to form a closed circuit. Therefore, the minimum voltage can be computed as follow.

$$\Delta V_{bs} = V_{cc} - V_f - V_{ge} - V_{ce} - V_{rs} \quad (7)$$

Where  $V_{ge}$  is the minimum voltage that can turn the power element at the low side during the charging process.

Besides,  $Q$  charge is defined by the intensity of the current flowing between the two plates in a unit of time. So, the  $Q$  charge total of the bootstrap capacitor is a sum of  $Q$  charge of each element. The total charge of bootstrap capacitors is presented in (8) [14], where the element charges include the gate charge of high side IGBT ( $Q_g$ ), level shift charge required per cycle ( $Q_{ls}$ ), and level shift charge required per cycle ( $Q_{qbs}$ ).

$$Q_{total} = 2Q_g + Q_{ls} + Q_{qbs} \quad (8)$$

The bootstrap capacitor calculated in (6) is the minimum capacitor value. In fact, due to manufacturing errors or external factors, selecting the capacitor with the minimum value can lead to overcharging, causing IC damage. To avoid this, it is common practice to select the capacitor is larger 2 times than the minimum value at least.

Not only the importance of the bootstrap capacitor, but also the resistor and diode bootstrap. In particular, a fast recovery diode should be used to limit the quantity of charge returning to the VCC when the bootstrap capacitor is released. Furthermore, a resistor must be connected immediately with the bootstrap diode in order to limit the charging current to the capacitor.

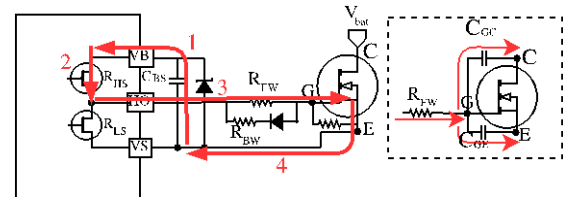


Fig. 8. The current path of the capacitor discharge process of high side

Additionally, the output gate resistance should be considered as it affects the power device's switching time. when the energy from the bootstrap capacitor is released and creates  $I_g$  current flows through the  $R_{HS}$ , which is internal resistance in the high side of the IC bootstrap, and then pass  $R_{FW}$  (the red-line path in figure 8). The result is the high side power device will be turned on. With the  $R_{HS}$  resistor value is constant during the operating time, the  $I_g$  current will be based on the  $R_{FW}$  value, so that the high side gate can be turned on at the lowest voltage (threshold voltage of G-E).

$$I_g = \frac{V_{bs} - V_{ge(th)}}{R_{HS} + R_{FW}} \quad (9)$$

$$R_{FW} = \frac{V_{bs} - V_{ge(th)}}{I_g} - R_{HS} \quad (10)$$

Moreover, there are internal capacitors between each terminal of IGBT [15]. So, when supplying voltage to G terminal, the current will be charged to the capacitor in G-E terminal first. In the other hand, the voltage at G-E terminal is proportional to the opening of the C-E terminal, so that the C-E voltage changes as a result in the change of voltage at G-E terminal, and the C-G terminal capacitor is started to charge, this is due to the Miller effect [16]. So, considering the amount of  $Q$  charge during the high side period is on, the current  $I_g$  is determined as follows.



$$I_g = \frac{Q_{GE} + Q_{GC}}{T_{ON}} \quad (11)$$

However, when either both the IGBT on the high side or the low side switches from the open state to the closed state, then there will be an energy at the collector-gate discharging backward [17]. And if this reverse voltage is greater than the threshold voltage gate-emitter that will cause IGBT to activate itself, the result leads to the voltage supply of power device immediately short to the ground, the circuit will be burnt out. To avoid this, a diode and a resistor should be added to recover the pulse and reduce the reverse voltage to ensure the reverse voltage less than the threshold voltage of the G-E terminal for no self-conductivity occurs.

$$I_g(R_{BW} + R_{LS}) + V_f < V_{CE(th)} \quad (12)$$

$$R_{BW} < \frac{V_{CE(th)} - V_f}{I_g} - R_{LS} \quad (13)$$

### B. Low-pass LC filter

Because the sine wave is created by converting IGBT power semiconductor devices, so the output waveform cannot avoid the voltage ripples and high-frequency ripples from power devices. Therefore, the use of an LC filter is necessary to eliminate high-order harmonics, help to minimize output ripples [18]. A low pass filter includes two elements that are an inductor and a capacitor, also known as a second-order filter, connected in series at each phase as shown in figure 1 above. There are two important parameters when designing LC filters that include the cut off frequency ( $f_c$ ), and the quality factor ( $Q$ ). The function of an LC filter is to attenuate waves with a higher frequency than the cut off frequency and to allow waves with a frequency smaller than the cut off frequency to pass through. Moreover, the selection of the cut off frequency directly affects the control bandwidth. In particular, increasing quality factor  $Q$  will reduce the control bandwidth. The reduction of bandwidth causing phase delays at the output due to the attenuation of speed response of the inverter system [19]. To compromise between attenuation and phase response, the  $Q$  factor should be around 0.707 following the principle of Butterworth [20]. At this value, it achieves the flatness at then the expense of a relatively wide transition region from passband to stopband (have no peaking) [21]. The values of inductor and capacitor should be selected accordingly, if the cutoff frequency selected is too low, it may attenuate the operating frequency range resulting in a decrease in output power. Conversely, if the cutoff frequency selected is too high, close to the switching frequency, the LC filter is almost ineffective. Typically, the cut off frequency must be approximately  $1/10^{\text{th}}$  of the conversion frequency (14) [22], and the quality coefficient formula  $Q$  is determined accordingly in (15) [23].

$$f_c = \frac{1}{2\pi\sqrt{LC}} = \frac{f_{sw}}{10} \quad (14)$$

$$Q = R_L \sqrt{\frac{C}{L}} \quad (15)$$

Where  $f_{sw}$  is the switching frequency and  $R_L$  is the load resistance. Considering the case of maximum load, the load resistance is equal to the square of RMS voltage divided by rated power [24].

From (14) - (15), the values of  $C$  and  $L$  are the roots of a system of two (16) and (17).

$$LC = \frac{25}{\pi^2 f_{sw}^2} \quad (16)$$

$$\frac{C}{L} = \frac{Q^2 P_{rated}}{V_{RMS}^4} \quad (17)$$

So,

$$L = \frac{5V_{RMS}^2}{\pi f_{sw} Q P_{rated}} \quad (18)$$

$$C = \frac{5Q P_{rated}}{\pi f_{sw} V_{RMS}^2} \quad (19)$$

### C. Shunt resistor for current sensing circuit

The shunt resistor plays the role of sensing the current to feedback to the microcontroller. Typically, there are two ways to layout the resistance, the first way is to connect resistance at each phase of the motor. This can be monitored each of the individual phases, but this requires the microcontroller's resources, requiring at least three microcontroller ADC pins for this monitoring. Moreover, the signal of the current-feedback may be heterogeneous due to the error of each resistor. The second way is to connect the resistor at the node of the three low side IGBTs, which only need to use one ADC pin to read the signal.

However, the feedback current to the microcontroller is very small, because the shunt resistor is usually very low, around a few  $m\Omega$  to avoid losing due to heat. Specifically, in Atmega328P microcontroller, the voltage of ADC pins is referenced at 5V with a resolution of 10 bits equal to 1024 values, so that the minimum division is approximately 4.9mV. Therefore, to read the signal of the microcontroller more accurately, the signal should be passed through a voltage amplifier using an IC Operation Amplifier (Op-amp) is shown in figure 9 below.

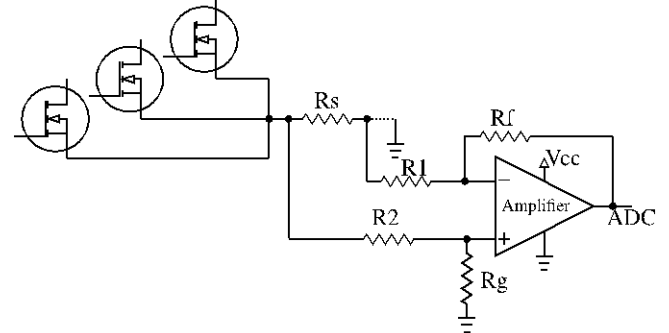


Fig. 9. The layout of shunt resistor

Op-amp is used as a differential amplifier, responsible for receiving the voltage signals from the shunt resistor then amplify this voltage signal and send it to the microcontroller. The amplification voltage is determined by multiply the voltage measured at the shunt resistor by the amplification factor, which is shown in (20) [25].

$$V_{ADC} = V^+ \left[ \frac{(R_f + R_I)R_g}{(R_g + R_2)R_I} \right] - V^- \left( \frac{R_f}{R_I} \right) \quad (20)$$

As can be seen from the figure 9, the reference voltage  $V^-$  is pulled to the ground while the voltage  $V^+$  is the current  $I$  through the resistor  $R_s$ . In addition, for simplicity, resistors  $R_I$  and  $R_g$  are chosen to be equal to  $R_2$  and  $R_f$ . Therefore, the  $V_{ADC}$  feedback can be abbreviated as follows.

$$V_{ADC} = IR_s \frac{R_f}{R_I} \quad (21)$$

## V. EXPERIMENTAL SETUP

In this driver circuit, the power semiconductor device and the IC bootstrap used in this experiment are IHW20N120R2 [26] and IR2103 [27] respectively. The calculation will be implemented based on the formulas presented above. Then, a prototype of a driver circuit is formed.

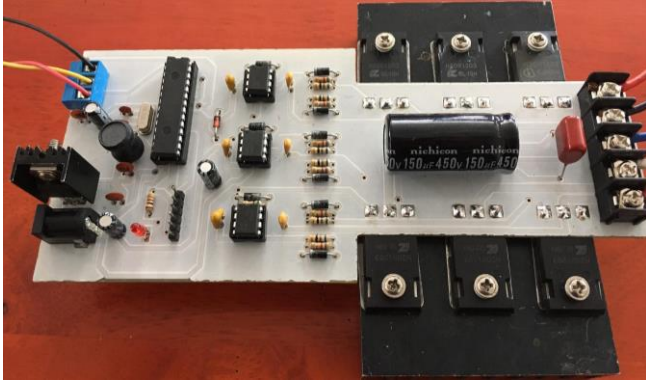


Fig. 10. A prototype of driver circuit

Figure 11 below shows the electric motorcycle model.



Fig. 11. The induction motor applied in electric motorcycle model.

The motor is connected to the rear wheel by a sprocket system, which is shown in figure 12.



Fig. 12. The powertrain of electric motorcycle model.

This driver circuit performs to drive a 2-HP squirrel cage induction motor with the specification shown in table I.

TABLE I. SPECIFICATION OF INDUCTION MOTOR

Parameter	Value	Unit
Nominal output	2.0	Hp
Frequency rated	50	Hz
Voltage rated		
Star connection	220	V
Delta connection	380	
Pole number	4	Pole
Nominal speed	1450	RPM

## VI. EXPERIMENTAL RESULTS

In this experiment, the driver circuit drives squirrel cage induction motor under approximately switching frequency 7800Hz based on (2) and the fundamental frequency at 50Hz (rated) to measure the output control signals. There are a total of 6 output control signals from the microcontroller that export to IR2103, divided into 3 pairs. Figure 13 below shows the output signal of a pair of two signals. In which, channel 1 is the SPWM signal that is transmitted to the HIN pin of IR2013 while channel 2 is the signal transmitted to the LIN pin.

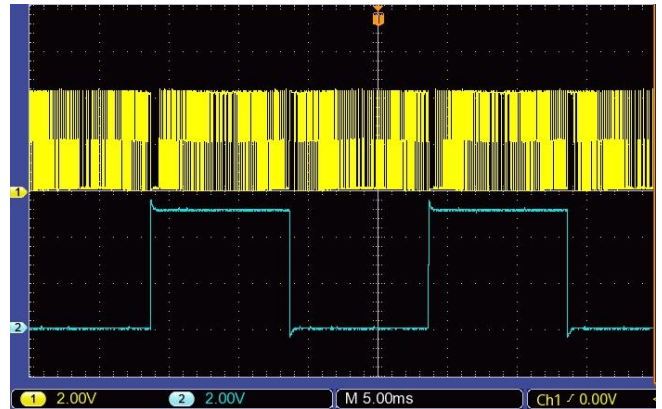


Fig. 13. The waveform of signals at HIN and LIN pins

Figure 14 shows the output voltage at HO and LO pins of IR2103 that use to control the power devices at the high side and low side. The power devices perform an alternating turn on and off continuously to create voltage interruption so that the overall voltage at the output of each phase is similar to the sinusoidal voltage.

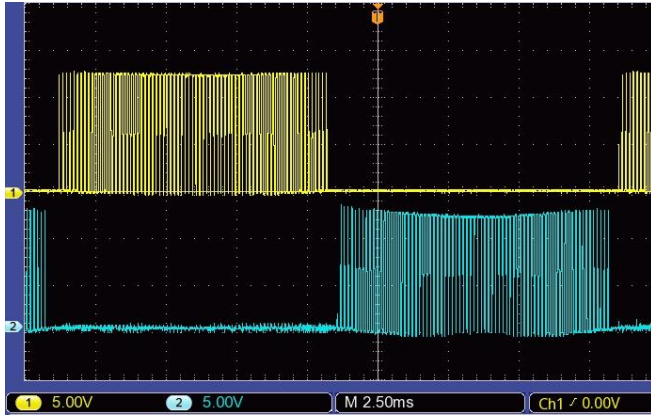


Fig. 14. The waveform of control signals at HO and LO of the gate driver

Figure 15 shows the output voltage waveform of a single-phase at fundamental frequency.

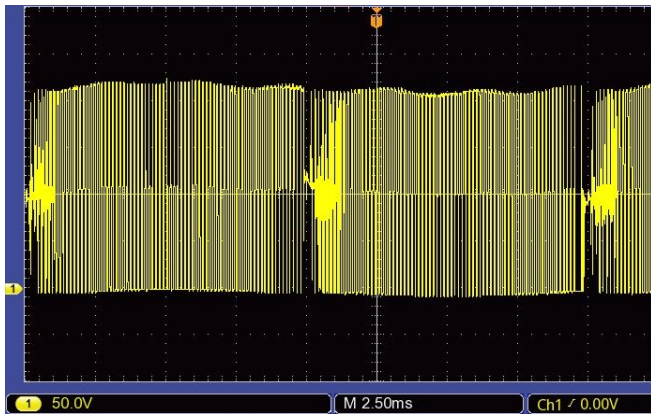


Fig. 15. The experimental waveform of output voltage

Figure 16 shows the waveform of the output current, which measured at a 5mΩ shunt resistor with an Op-amp output amplification factor of 65.

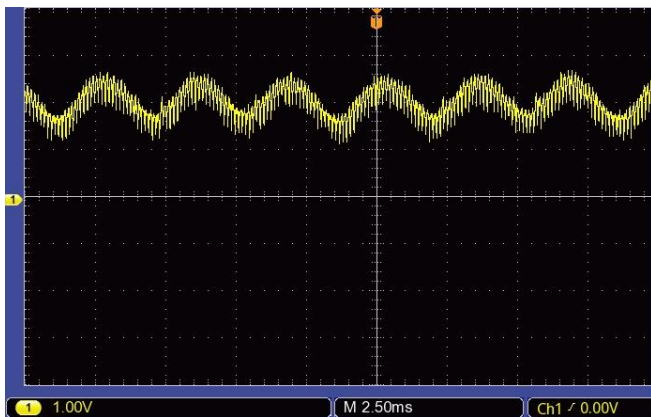


Fig. 16. The experimental waveform of current feedback at 6A

Figures 17 and 18 show the motor speed in two cases of no-load and 2.5Nm load respectively at 30Hz for 10 seconds.

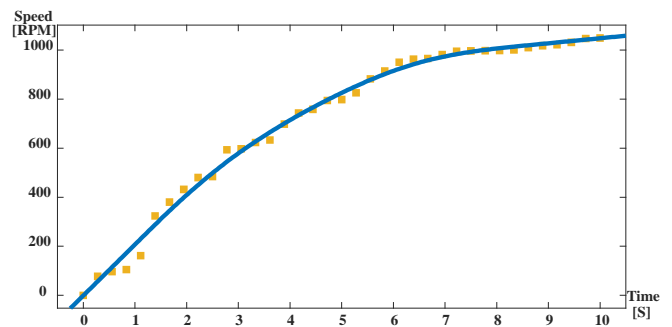


Fig. 17. The experimental speed-time curve during run up at no load

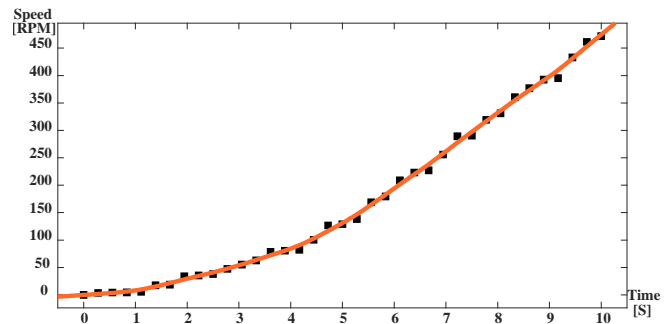


Fig. 18. The experimental speed-time curve during run up at a constant load

## VII. CONCLUSION

The main purpose of this research is to design a Vietnamese electric motorcycle model that encourage people to ride more electric motorcycles to reduce emissions affecting the environment.

In that spirit, the first step that needs to be taken is to design the driver circuit, an indispensable component of the induction motor. This paper presents the control principles based on the working principles of VFD. The theoretical basis and computations are then built to select the appropriate electronic components in this circuit. Finally, the experiment was practiced on a 2 HP squirrel cage motor that applied on an electric motorcycle to perform the necessary assessments.

## REFERENCES

- [1] C. Chan, "The state of the art of electric and hybrid electric," in Proceedings of the IEEE, vol. 90, n° 2, February 2002.
- [2] Carl Vogel, "Build Your Own Electric Motorcycle," the McGraw-Hill Companies, July 2009.
- [3] Hoang, Tuan Anh, Nam Chu and Trung V. Tran, "The Environmental Pollution In Vietnam: Source, Impact And Remedies," International Journal of Scientific & Technology Research 6 (2017): 249-253.
- [4] Jamadar, Kumbhar, Gavane, Sutrave, "Design and Development of Control System for Three Phase Induction Motor using PIC Microcontroller," IFAC Proceedings Volumes (IFAC PapersOnline). 3. 807-811. 10.3182/20140313-3-IN-3024.00072, March 2014.
- [5] Hossein Madadi Kojabadi, "A comparative analysis of different pulse width modulation methods for low cost induction motor drives," in Energy Conversion and Manage, Volume 52, Issue 1, pages 136-146, January 2011.
- [6] X. Liang, R. Laughy and J. Liu, "Investigation of Induction Motors Starting and Operation with Variable Frequency Drives," 2007

- Canadian Conference on Electrical and Computer Engineering*, Vancouver, BC, pp. 556-561, doi: 10.1109/CCECE.2007.144, 2007.
- [7] Rakesh Parekh, "AC Induction Motor Fundamentals," in DS00887A, Microchip Technology Inc., 2003.
  - [8] Gopal K. Dubey, "Fundamentals of Electrical Drives," 2<sup>nd</sup> Edition, Narosa Publishing House, New Delhi, 2011.
  - [9] Z. Zhang, Y. Liu and A. M. Bazzi, "An improved high-performance open-loop V/f control method for induction machines," 2017 IEEE Applied Power Electronics Conference and Exposition (APEC), Tampa, FL, 2017, pp. 615-619, doi: 10.1109/APEC.2017.7930757.
  - [10] Austin Hughes, "Electric motor and Drives," 3<sup>rd</sup> Edition, Elsevier, December 2005.
  - [11] Atmel Corporation, "ATmega328P," Rev: 7810D-AVR-01/15, 2015.
  - [12] Ronald N. Jansen, Steven T. Haensgen, "Method and apparatus for calculating RMS value," in Rockwell Automation Technologies, Inc., Mayfield Heights, OH (US), February 2003.
  - [13] Mitsubishi Electric, "DIPIPM application note bootstrap circuit design manual" in Dual-in-Line Package Intelligent Power Module, Tokyo, Japan: Mitsubishi Electronic Corp, 2016.
  - [14] J. Adams, "Bootstrap component selection for control ICs," Proc. DT-98-2a Int. Rectifier, 2001.
  - [15] Vinod Kumar Khanna, "Insulated Gate Bipolar Transistor IGBT Theory and Design," IEEE Press-Wiley Interscience, New Jersey, USA, August 2003.
  - [16] J. Boehmer, J. Schumann and H. Eckel, "Effect of the miller-capacitance during switching transients of IGBT and MOSFET," 15th International Power Electronics and Motion Control Conference (EPE/PEMC), Novi Sad, 2012, pp. LS6d.3-1-LS6d.3-5, 2012.
  - [17] Fuji Electric, "Fuji IGBT Modules Application Manual," in Innovating Energy Technology, March 2015.
  - [18] K. H. Ahmed, S. J. Finney and B. W. Williams, "Passive Filter Design for Three-Phase Inverter Interfacing in Distributed Generation," Compatibility in Power Electronics, Gdansk, pp. 1-9, 2007.
  - [19] Hojabri, Mojgan, "Design, application and comparison of passive filters for three-phase grid-connected renewable energy systems," ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 22, December 2015.
  - [20] George Ellis, "Control System Design Guide," 4<sup>th</sup> Edition Elsevier, May 2012.
  - [21] Analog Devices Inc. Engineeri, "Linear Circuit Design Handbook," 1st Edition, Elsevier, February 2008.
  - [22] Kim, Hyosung and Sul, Seung-Ki, "A Novel Filter Design for Output LC Filters of PWM Inverters," Journal of Power Electronics, vol. 11, no. 1, pp. 74-81, Jan 2011.
  - [23] Texas Instrument Incorporated, "LC Filter Design," in SLAA701A Application Report, November 2016.
  - [24] Ahmad Ale Ahmd, "A New Design Procedure for Output LC Filter of Single Phase Inverters," in International Conference on Power Electronics and Intelligent Transportation System, At China, Volume: 3rd, 2010.
  - [25] Ron Mancini, "Op Amps For Everyone," in Advanced Analog Products, Texas Instrument, August 2002.
  - [26] Infineon Technologies AG, "IHW20N120R2," in Power Semiconductors, Rev. 1.2, July 2006.
  - [27] International Rectifier, "IR2103(s)PBF Half-Bridge Driver," April 2013.

# Fractional order Modeling and Control of a Quadruple-tank Process

Lam Chuong Vo  
Mechanical Engineering Department  
University of Technology and  
Education  
Ho Chi Minh City, Viet Nam  
chuongvl@hcmute.edu.vn

Luan Vu Truong Nguyen  
Mechanical Engineering Department  
University of Technology and  
Education  
Ho Chi Minh City, Viet Nam  
vluuantn@hcmute.edu.vn

Moonyong Lee  
School of Chemical Engineering  
Yeungnam University  
South Korea  
mynlee@yumail.ac.kr

**Abstract**— In recent years, fractional calculus attracts considerable attention from researchers in modeling and control of systems due to its flexibility in describing the behaviors of the dynamic systems. In this paper, a control design method based on fractional order will be proposed for a multi-input multi-output (MIMO) process. To deal with the interactions between system variables, the decoupling technique using simplified decoupling is adopted. However, in this paper, the PSO algorithm is proposed to approximate the complex decoupled transfer function into a simple form of fractional-order transfer function. The tuning rules of a fractional-order PID (FOPID) controller are also derived analytically based on the internal model control (IMC) structure. To validate the proposed method in real applications, the identification of a multivariable process (MIMO), the quadruple tank system, is presented by using the diagonal form of matrix fraction description (MFD) which transforms a MIMO process into multi-input single-output (MISO) sub-models. Then, the proposed controller is simulated as well as implemented based on the identified models to illustrate the effectiveness of the proposed method.

**Keywords**— Fractional order controllers, PSO algorithm, simplified decoupling, Quadruple-tank system.

## I. INTRODUCTION

Multi-input multi-output (MIMO) systems consist of many process variables including measurement and control signals with complicated interactions that make it difficult to control. The well-known quadruple-tank process, proposed by Johansson [1], is a common system for evaluating control strategies of multivariable processes. Many researchers have developed control algorithms for improving its performances in terms of servomechanism problem and disturbance rejection [2–7]. However, from the conventional controllers to the advanced control algorithms, the process delays are neglected in the system model. In real applications, they are responsible for the degradation of system performance at higher values of gain [7].

In this work, the simplified decoupling method proposed by Vu et al. [8] is adopted to deal with interactions between process variables. To overcome the realizability problem of the decoupling techniques and enhance dynamic behaviors of decoupled systems, fractional-order processes are suggested to be the equivalent transfer functions of the decoupled elements. The particle swarm optimization (PSO) algorithm for the approximation procedure proposed in [9] is employed to find out the parameters of approximated fractional functions. Bouyedda et al. [10] performed a similar work by using the Genetic Algorithm (GA) to reduce a high integer-order transfer function of a SISO system to a lower fractional-order one.

Fractional calculus is a mathematical phenomenon that helps to describe dynamic behaviors of real plants more accurately than the integer-order conventional methods. In the control field, fractional calculus had been widely used since Oustaloup introduced an approximation approach of fractional derivative and integral in the frequency domain [11]. Therefore, in recent years, the fractional-order proportional-integral-derivative (FOPID) controller has attracted more attention from many researchers. The FOPID has five tuning parameters including proportional, integral, derivative gain, and fractional orders of the integral and derivative terms which provide more flexibility in system performances as well as robustness compared with the conventional ones [12–14]. Because of more tuning parameters, it is also harder to derive analytical tuning rules for the controller. Different tuning methods have been suggested to solve this kind of problem [15–20]. However, most of them are used to deal with single-input single-output (SISO) systems.

In this paper, the FOPI based IMC is presented to find out analytical tuning rules of the controllers for MIMO processes and it is also validated on the quadruple-tank system. The time delays of the system are also considered in the design procedure by using the Padé approximation technique for exponential terms [25]. The proposed method uses the IMC scheme to reduce the number of tuning parameters; and normally, there is only one parameter left needs to be tuned based on some criteria of system performances in terms of set-point tracking as well as disturbance rejection [10, 18–20]. The identification of process parameters plays a crucial role in multivariable control design for industrial applications because a better performance will be achieved by model-based tuning algorithms. In this work, the least-squares method which is reliable technique for system identification is used for MIMO systems. The diagonal form of matrix fraction description is suggested to transform a MIMO system into several multi-input single-output sub-models [21–23].

This paper is organized as follows. Section 2 is briefly introduced fractional-order calculus with the Oustaloup recursive algorithm to approximate the fractional operator. Section 3 will discuss the Quadruple tank system including its dynamics and the approach used for system identification. The controller structure is mentioned in section 4; and a new fractional-order PI controller based on the IMC structure is also proposed and its analytical tuning rules are derived in this section. The obtained controller will be justified on the real model and the results are included in section 5. Finally, conclusions are given in section 6.



## II. PRELIMINARY

### A. Fractional order calculus

Fractional calculus is a generalization of ordinary calculus by extending the integration and differentiation order to the non-integer order. It presented a fractional operator  ${}_a D_t^\nu$  where  $a$  and  $t$  are the limits and  $\nu$  is the fractional order ( $\nu \in \mathbb{R}$ ). The most common definition of the fractional operator was proposed by Riemann and Liouville [13] and it is defined as following

$${}_a D_t^\nu = \frac{1}{\Gamma(n-\nu)} \frac{d}{dt} \int_a^t \frac{f(\tau)}{(t-\tau)^{\nu-n+1}} d\tau, n-1 < \nu < n \quad (1)$$

where  $\Gamma(\bullet)$  represents the Euler's gamma function; with a positive  $\nu$ , the fractional operator denotes fractional derivative, and a negative  $\nu$  represents fractional integral.

The fractional-order transfer function of a SISO system can be described:

$$G(s) = \frac{b_m s^{\lambda_m} + b_{m-1} s^{\lambda_{m-1}} + \dots + b_0 s^{\lambda_0}}{a_n s^{\nu_n} + a_{n-1} s^{\nu_{n-1}} + \dots + a_0 s^{\nu_0}} \quad (2)$$

The fractional-order of  $s$  in equation (2) makes it difficult to simulate or implement a fractional-order system. Therefore, Oustaloup proposed an approximation method using a recursive algorithm with finite numbers of poles and zeros for most applications [11, 13]. Within the specific range of frequency  $[\omega_b, \omega_h]$ , the approximation of the FO operator,  $s^\nu$ , can be obtained by the following equation:

$$s^\nu \cong s_{[\omega_b, \omega_h]}^\nu \simeq K \sum_{k=-N}^N \frac{s + \omega_z}{s + \omega_p} \quad (3)$$

where the zero, pole and gain can be calculated respectively from:

$$K = \omega_h^\nu \quad (4)$$

$$\omega_z = \omega_b \left( \frac{\omega_h}{\omega_b} \right)^{(k+N+0.5-0.5\nu)/(2N+1)} \quad (5)$$

$$\omega_p = \omega_b \left( \frac{\omega_h}{\omega_b} \right)^{(k+N+0.5+0.5\nu)/(2N+1)} \quad (6)$$

## III. QUADRUPLE-TANK SYSTEM

### A. The system description and its dynamics

Fig. 1 shows the pipe and instrumentation diagram (P&ID) of the quadruple-tank system. The inlet liquids are pumped by two centrifugal pumps into four tanks with cross-connection as shown in the figure. Manipulated variables are two analog voltages ( $u_1, u_2$ ; 0÷10 VDC) which applied to the inverters to control the power of the two pumps and then manipulate the flow rates of the inlet streams. The two three-way valves ( $V_1, V_2$ ) with the adjustable coefficients  $\gamma_1, \gamma_2 \in [0, 1]$  determine the portion of their outputs to the upper and lower tanks respectively. The outputs of this system are two levels of the lower tanks ( $h_1, h_2$ ) which are measured by level sensors, LR2750 of IFM (LT<sub>1</sub>, LT<sub>2</sub>)

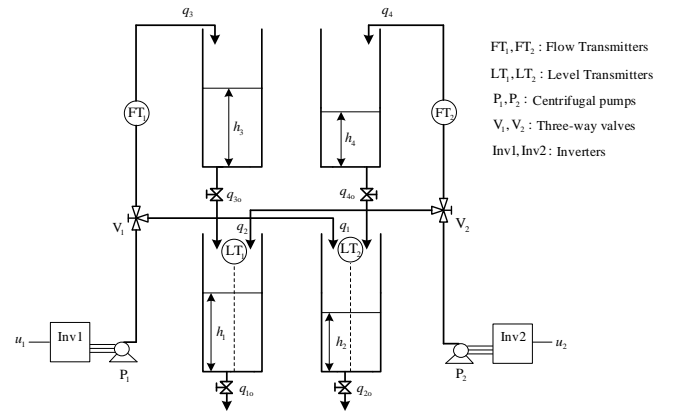


Fig 1. The piping and instrumentation diagram of the quadruple-tank system

Applying the mass conservation law and Bernoulli's law, the dynamic equations of the system are available in some literature [1, 4, 6, 7]. In this work, they are linearized around the operating points:  $h_i^0$  corresponding to the inputs ( $u_1^0, u_2^0$ ). And let define new variables:

$$x_i = h_i - h_i^0 \quad (i = 1 \div 4) \quad (7)$$

$$r_j = u_j - u_j^0 \quad (j = 1 \div 2) \quad (8)$$

Then, the linearization equations are obtained as following:

$$\frac{dx_1}{dt} = -\frac{1}{\tau_1} x_1 + \frac{A_3}{A_1 \tau_3} x_3 + \frac{\gamma_2 k_2 r_2}{A_1} \quad (9)$$

$$\frac{dx_2}{dt} = -\frac{1}{\tau_2} x_2 + \frac{A_4}{A_2 \tau_4} x_4 + \frac{\gamma_1 k_1 r_1}{A_2} \quad (10)$$

$$\frac{dx_3}{dt} = -\frac{1}{\tau_3} x_3 + \frac{(1-\gamma_1) k_1 r_1}{A_3} \quad (11)$$

$$\frac{dx_4}{dt} = -\frac{1}{\tau_4} x_4 + \frac{(1-\gamma_2) k_2 r_2}{A_4} \quad (12)$$

$$\text{where } \tau_i = \frac{A_i}{a_i} \sqrt{\frac{2h_i^0}{g}} \quad (i = 1 \div 4): \text{time constant (s)} \quad (13)$$

$A_i$  is the cross-sectional area of the tank  $i$ ,  $a_i$  is the cross-sectional area of the outlet hole of the tank  $i$ ; it is assumed that the characteristics of the pumps are linear. So, the relationship between the applied voltage and the outlet flow rate of each pump is a constant represented by  $k_i$ ; it is also ignored the delay time for the liquid that travels from the pump to the tanks.

Using Laplace transform to convert the differential equations into the matrix form of transfer functions, we obtain:

$$\begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix} = \begin{bmatrix} \frac{K_1(1-\gamma_1)}{(\tau_1 s + 1)(\tau_3 s + 1)} & \frac{K_2 \gamma_2}{\tau_1 s + 1} \\ \frac{K_3 \gamma_1}{\tau_2 s + 1} & \frac{K_4(1-\gamma_2)}{(\tau_2 s + 1)(\tau_4 s + 1)} \end{bmatrix} \begin{bmatrix} R_1(s) \\ R_2(s) \end{bmatrix} \quad (14)$$

$$\text{where } K_1 = \frac{\tau_1 k_1}{A_1}, K_2 = \frac{\tau_1 k_2}{A_2}, K_3 = \frac{\tau_2 k_1}{A_2}, K_4 = \frac{\tau_2 k_2}{A_2}$$

Note that, in Eq. (14), for its simplicity, the time delays are ignored. In real-time application, however, the liquid need time to travel from the pumps to the tanks. Therefore, each

transfer function in Eq. (14) will be in series with a time delay term.

### B. Identification of the Quadruple-tank system

The theoretical model of the quadruple-tank system is derived in the previous section. It is shown that the system nonlinear model could be approximated to the linear form of the two-input, two-output (TITO) system. In this part, the least-squares method will be adopted to identify each transfer function of the whole system. The general linear form can be described by:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} G_{11}(q) & G_{12}(q) \\ G_{21}(q) & G_{22}(q) \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} + \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} \quad (15)$$

where  $G_{ij}(q)$  is a rational of polynomials in the delay operator  $q^{-1}$

In this paper, the diagonal form of matrix fraction description (MFD) is used due to its simplicity and applicability. As a result of that, the TITO process is decoupled into two two-input single-output sub-models which are identified separately. The diagonal form MFD is given as

$$\mathbf{G}(q) = \begin{bmatrix} A_1(q) & 0 \\ 0 & A_2(q) \end{bmatrix}^{-1} \begin{bmatrix} B_{11}(q) & B_{12}(q) \\ B_{21}(q) & B_{22}(q) \end{bmatrix} \quad (16)$$

where  $A_i(q)$  and  $B_{ij}(q)$  ( $i, j = 1 \div 2$ ) are polynomials in the delay operator  $q^{-1}$ :

$$A_i(q) = 1 + a_{i,1}q^{-1} + a_{i,2}q^{-2} + \dots + a_{i,na_i}q^{-na_i} \quad (17)$$

$$B_{1j}(q) = b_{1j,1}q^{-1} + b_{1j,2}q^{-2} + \dots + b_{1j,nb_{1j}}q^{-nb_{1j}} \quad (18)$$

$$B_{2j}(q) = b_{2j,1}q^{-1} + b_{2j,2}q^{-2} + \dots + b_{2j,nb_{2j}}q^{-nb_{2j}} \quad (19)$$

Replace (16) into (15), it yields:

$$\begin{cases} A_1(q)y_1(t) = B_{11}(q)u_1(t) + B_{12}(q)u_2(t) + A_1(q)v_1(t) \\ A_2(q)y_2(t) = B_{21}(q)u_1(t) + B_{22}(q)u_2(t) + A_2(q)v_2(t) \end{cases} \quad (20)$$

Considering the first sub-model, the least-squares estimation of the system parameters to minimize the loss function:

$$V_{LS1} = \sum_{t=na_1+1}^N (A_1(q)y_1(t) - [B_{11}(q)u_1(t) + B_{12}(q)u_2(t)])^2 \quad (21)$$

The solution is obtained in the following form [21, 23]

$$\hat{\Phi} = [\Phi_1^T \Phi_1]^{-1} \Phi_1^T \mathbf{y}_1 \quad (22)$$

where

$$\mathbf{y}_1 = \begin{bmatrix} y_1(na_1 + 1) \\ y_1(na_1 + 2) \\ \vdots \\ y_1(N) \end{bmatrix};$$

$$\Phi = \begin{bmatrix} a_{1,1} & \dots & a_{1,na_1} & b_{11,1} & \dots & b_{11,nb_{11}} & b_{12,1} & \dots & b_{12,nb_{12}} \end{bmatrix}^T$$

$$\Phi_1 = \begin{bmatrix} y_1(na_1) & \dots & y_1(1) & u_1(na_1) & \dots & u_2(1) \\ \vdots & & \vdots & \vdots & & \vdots \\ y_1(N-1) & \dots & y_1(N-na_1) & u_1(N-1) & \dots & u_2(N-na_1) \end{bmatrix}$$

Using the PRBS signal to excite the system around its initial operating point, then the pair of input-output data are collected. The sampling time is chosen as  $T_s = 0.1$  (s) and after the pretreatment of the data, 2000 samples are available. The first 1200 samples are used for model identification and the rest of data (800 samples) are used for validation. Fig. (2) and (3) prove that the identified models are quite accurate, and therefore, can be used for control purposes.

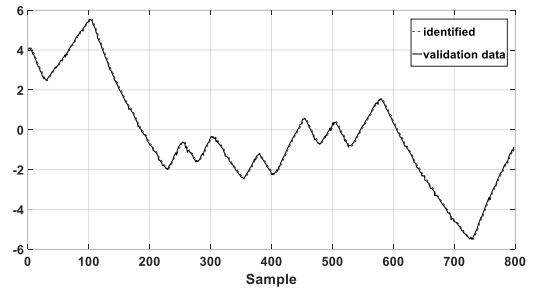


Fig 2. The validation of identified sub-model 1

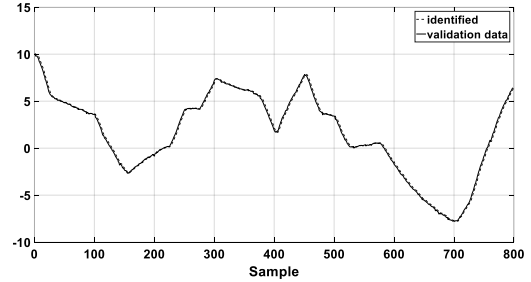


Fig 3. The validation of identified sub-model 2

In both sub-models, using  $na_i = 2$ ;  $nb_{1j} = 2$ ;  $nb_{2j} = 2$ , the model parameters are obtained as follows:

$$\Phi_1 = [-1.3692 \quad 0.3725 \quad 0.0743 \quad 0.0737]^T \quad (23)$$

$$\Phi_2 = [-1.2289 \quad 0.2305 \quad 0.0023 \quad 0.0722]^T \quad (24)$$

Convert into the continuous transfer functions with the specified sampling time. The elements of the matrix transfer function of the system are derived and approximated by first-order plus time-delay systems. The results are summarized in Table 1.

TABLE I. THE IDENTIFIED TRANSFER FUNCTIONS AND THEIR APPROXIMATION

Identified transfer functions	Approximated
$G_{11}(s) = \frac{-0.8229s+15.08}{s^2+12.46s+0.6891}$	$\bar{G}_{11}(s) = \frac{21.885e^{-0.1s}}{18.0364s+1}$
$G_{12}(s) = \frac{-0.7136s+15.2}{s^2+12.46s+0.6891}$	$\bar{G}_{12}(s) = \frac{22.0306e^{-0.8s}}{17.2757s+1}$
$G_{21}(s) = \frac{-0.2244s+0.7376}{s^2+20.53s+0.4869}$	$\bar{G}_{21}(s) = \frac{1.5101e^{-2s}}{39.9664s+1}$
$G_{22}(s) = \frac{-1.694s+20.13}{s^2+20.53s+0.4869}$	$\bar{G}_{21}(s) = \frac{41.2358e^{-1.7s}}{40.2471s+1}$

#### IV. SIMPLIFIED DECOUPLING BASED ON FRACTIONAL-ORDER SYSTEMS

##### A. Simplified decoupling based on fractional-order systems

In this work, the simplified decoupling is used to deal with the main issue of multivariable processes, the interactions between process variables. The control structure is shown in Fig 4 where  $\mathbf{G}(s)$ ,  $\mathbf{D}(s)$ ,  $\mathbf{Q}(s)$  are the transfer function matrix, the decoupling matrix, and the decoupled matrix of a TITO process, respectively;  $\mathbf{G}_c(s) = \text{diag}(G_{c1}(s), G_{c2}(s))$  is the diagonal form of the controller matrix. Due to the properties of the simplified decoupling [8, 9], they have the forms as following:

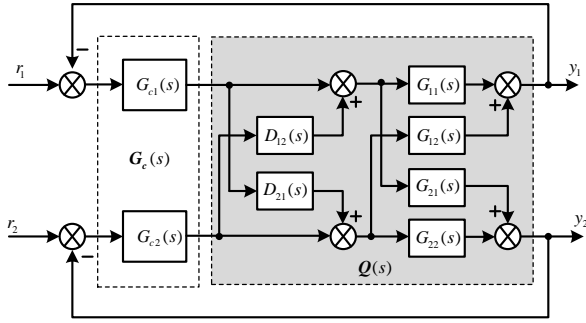


Fig 4. The simplified decoupling structure with fractional control

$$\mathbf{G}(s) = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix}; \mathbf{D}(s) = \begin{bmatrix} 1 & D_{12}(s) \\ D_{21}(s) & 1 \end{bmatrix};$$

$$\mathbf{Q}(s) = \begin{bmatrix} Q_{11}(s) & 0 \\ 0 & Q_{22}(s) \end{bmatrix} \quad (25)$$

where the elements of  $\mathbf{D}(s)$  and  $\mathbf{Q}(s)$  can be calculated as follows [8, 9]:

$$D_{12}(s) = -\frac{\bar{G}_{12}(s)}{\bar{G}_{11}(s)} = -\frac{1.0067(18.0364s+1)}{17.2757s+1} \quad (26)$$

$$D_{21}(s) = -\frac{\bar{G}_{21}(s)}{\bar{G}_{22}(s)} = -\frac{0.0366(40.2471s+1)}{39.9664s+1} \quad (27)$$

$$Q_{11}(s) = \bar{G}_{11}(s) - \frac{\bar{G}_{12}(s)\bar{G}_{21}(s)}{\bar{G}_{22}(s)} = \frac{21.885e^{-0.1s}}{18.0364s+1} - \frac{22.0306e^{-0.8s} \cdot 1.5101e^{-2s}}{17.2757s+1} - \frac{39.9664s+1}{41.2358e^{-1.7s}} \quad (28)$$

$$Q_{22}(s) = \bar{G}_{22}(s) - \frac{\bar{G}_{12}(s)\bar{G}_{21}(s)}{\bar{G}_{11}(s)} = \frac{41.2358e^{-1.7s}}{40.2471s+1} - \frac{22.0306e^{-0.8s} \cdot 1.5101e^{-2s}}{17.2757s+1} - \frac{39.9664s+1}{21.885e^{-0.1s}} \quad (29)$$

The diagonal elements of the decoupled matrix are complicated as shown in Eq. (28)-(29). Therefore, various methods were proposed to approximate them to some standard forms. However, most of them only deal with integer-order transfer functions. In this paper, a fractional-order transfer function (FOTF) is proposed to be the equivalent transfer function of Eq. (28) and (29). The general form of the FOTF is as follow:

$$G_m(s) = \frac{K e^{-\theta s}}{\tau_2 s^{\alpha_2} + \tau_1 s^{\alpha_1} + 1} \quad (0 < \alpha_1 \leq 1 < \alpha_2 < 2) \quad (30)$$

where  $\tau_1, \tau_2$  are time constants;  $K$  is a gain;  $\alpha_1, \alpha_2$  are fractional orders;  $\theta$  is a delay time. In a special case, when  $\tau_2 = 0$ , Eq. (30) becomes the fractional first order transfer function.

The PSO algorithm for approximation proposed by Chuong and *et al.* [9] is adopted and expanded for the fractional-order case. Note that the parameter  $\theta$  is a prior value determined by the unit step response of the original model. Consequently, the number of tuning parameters, in this case, has the following vector form:

$$\mathbf{x} = [K \quad \tau_2 \quad \tau_1 \quad \alpha_2 \quad \alpha_1] \quad (31)$$

The different constraints of these parameters are given in (32):

$$\begin{cases} K_{min} < K < K_{max} \\ 0 < \tau_1 < \tau_{max} \\ 0 \leq \tau_2 < \tau_{max} \\ 0 < \alpha_1 \leq 1 \\ 1 < \alpha_2 < 2 \end{cases} \quad (32)$$

where  $K_{min}, K_{max}, \tau_{max}$  are determined based on the open-loop unit step response of the original model.

The results of the algorithm are shown in Eq. (33). Then, the fractional PI controller will be designed for each loop of the decoupled system.

$$\bar{Q}_{11}(s) = \frac{21.3233e^{-1.5s}}{15.8183s^{0.9523}+1} \quad \bar{Q}_{22}(s) = \frac{39.9728e^{-3s}}{40.2964s^{0.9728}+1} \quad (33)$$

##### B. Fractional IMC-PI controller design

In this study, a new structure of a fractional PI controller is proposed for each loop, called I<sup>o</sup>PI <sup>$\lambda$</sup> , and a lead/lag filter is also employed to improve the output performances. Consequently, the primary controller of each loop is as Eq. (34)

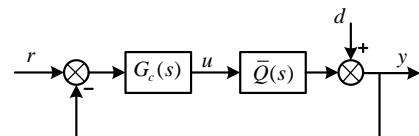
$$G_c(s) = K_c \frac{1}{s^\sigma} \left( 1 + \frac{1}{\tau_I s^\lambda} \right) F(s) \quad (34)$$

where  $K_c$  and  $\tau_I$  are proportional gain and integral time respectively;  $\lambda$  is fractional order of the integral term;  $\sigma$  is the fractional order of the ideal integral and  $\sigma = 1 - \lambda$ ; in special case, when  $\lambda = 1$  (integral term with integer order) then  $\sigma$  equals to zero (it becomes a conventional PI controller);  $F(s)$  is the lead-lag filter as Eq. (35)

$$F(s) = \frac{\tau_a s + 1}{\tau_b s + 1} \quad (35)$$

where  $\tau_a, \tau_b$  are time constants

The IMC based PID procedure normally used for integer-order processes is also addressed to design the proposed controller for fractional-order processes. Fig. 5 (a) and (b) show block diagrams of feedback control strategies including the classical feedback control and the internal model control as well [9, 18-20].



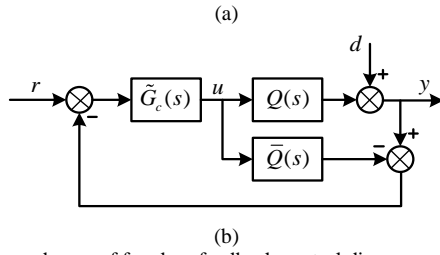


Fig. 5. The one degree of freedom feedback control diagram  
(a). Classical feedback control (b). Internal model control

In this work, the decoupled process model has the fractional first order form after approximating:

$$\bar{Q}(s) = \frac{K e^{-\theta s}}{\tau s^{\alpha+1}} \quad (0 < \alpha < 1) \quad (36)$$

The conventional controller is derived based on IMC structure as follow:

$$G_c(s) = \frac{\tau s^{\alpha+1}}{K[(\tau_c s + 1) - e^{-\theta s}]} \quad (37)$$

where  $\tau_c$  is an adjustable parameter which controls the tradeoff between the performance and robustness.

It is necessary to handle the time delay term in Eq. (37) properly to convert  $G_c$  into a familiar form of controller in industrial applications. In this paper, a Padé 1/1 approximation is used [25]:

$$e^{-\theta s} = \frac{1 - 0.5\theta s}{1 + 0.5\theta s} \quad (38)$$

The ideal feedback controller is obtained by:

$$G_c(s) = \frac{(\tau s^{\alpha+1})(1 + 0.5\theta s)}{K[(\tau_c s + 1)(1 + 0.5\theta s) - (1 - 0.5\theta s)]} = \frac{(\tau s^{\alpha+1})(1 + 0.5\theta s)}{Ks[0.5\tau_c\theta s + (\tau_c + \theta)]} \quad (39)$$

Rewritten Eq. (39) into the form of Eq. (34), the controller is derived as Eq. (40):

$$G_c(s) = \frac{\tau}{K(\tau_c + \theta)} \frac{1}{s^{1-\alpha}} \left( 1 + \frac{1}{\tau s^{\alpha}} \right) \frac{(0.5\theta s + 1)}{\frac{0.5\tau_c\theta}{\tau_c + \theta} s + 1} \quad (40)$$

$$K_c = \frac{\tau}{K(\tau_c + \theta)}; \tau_I = \tau; \lambda = \alpha;$$

$$\sigma = 1 - \alpha; \tau_a = 0.5\theta; \tau_b = \frac{0.5\tau_c\theta}{\tau_c + \theta}$$

## V. EXPERIMENT RESULTS

The realization of fractional control and the simplified decoupling technique is justified by an experiment on the real model. The obtained decoupler (Eq. 36 and 27) and the two diagonal elements of the decoupled matrix (Eq. 33) are used for controller design. The proposed fractional PI controllers are derived based on Eq. (40) in which  $\tau_c$  is chosen as 2 and 5 for each control loop respectively. The results are shown as follows:

$$G_{c1}(s) = 0.212 \frac{1}{s^{0.0477}} \left( 1 + \frac{1}{15.8183s} \right) \frac{0.75s+1}{0.4286s+1} \quad (41)$$

$$G_{c2}(s) = 0.126 \frac{1}{s^{0.0272}} \left( 1 + \frac{1}{40.2964s} \right) \frac{1.5s+1}{0.9375s+1} \quad (42)$$

The control structure is implemented by using Simulink of Matlab in Real-Time Window Target based on a supported

PCI card, 6323e of National Instrument. The time responses of the levels in both tanks are shown in Fig. 6 and Fig. 7.

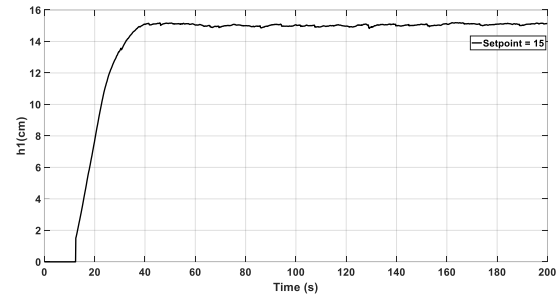


Fig 6. The step response of the level in tank1 with setpoint 15 (cm)

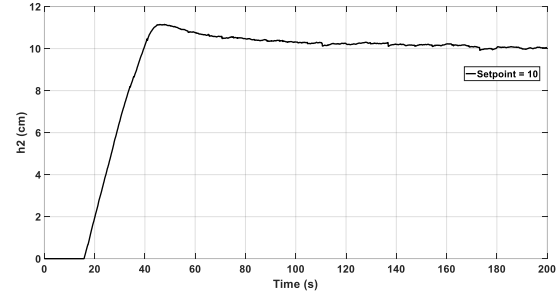


Fig 7. The step response of the level in tank2 with setpoint 10 (cm)

From the figures, it is obvious that the responses in both tanks are quite good. It is a fast response with no overshoot in tank 1. In tank 2, there is an overshoot of 10% and therefore, it makes the settling time is slower compared to that of tank 1. The steady-state errors of both tanks are about to zero.

## VI. CONCLUSIONS

In this paper, an approach to control a MIMO process is proposed by adopting the simplified decoupling technique and fractional PI controller. The PSO algorithm is also used to approximate the complex elements of the decoupled matrix to the fractional-order transfer function. The analytical tuning rules of the proposed fractional PI controller are derived based on the IMC structure. The tuning parameter is only a time constant to compromise the tradeoff between the system performance and its robustness. The Quadruple tank system is chosen to justify the effectiveness of the proposed method. A well-known identification algorithm, the least-squared method, is applied for a MIMO system to obtain the linearized model of the tank system. The experimental results prove the effectiveness of the proposed control structure in terms of setpoint tracking.

## ACKNOWLEDGMENT

This research was supported by the Science Research Program through Minister of Education and Training (MOET), as well as that of Ho Chi Minh City University of Technology and Education, Vietnam.

## REFERENCES

- [1] Johansson K.H.: The quadruple-tank process: a multivariable laboratory process with an adjustable zero. *IEEE Trans. Control System Technology* 8(3), 456–465 (1999).
- [2] Husek P.: Decentralized PI controller design based on phase margin specifications. *IEEE Trans. Control System Technology* 22(1), 346–351 (2014).
- [3] Panda C., Sujatha V.: Identification and control of multivariable system-role of relay feedback, *Introduction to PID Controllers-Theory, Tuning and Application to Frontier Areas*, InTech. (2012)

- [4] Ionescu C., Maxim A., Copot C., Kyeser R.: Robust PID auto-tuning for the quadruple tank system. 11<sup>th</sup> IFAC Symposium on Dynamics and Control of Process System including Biosystems, 919–924 (2016).
- [5] Basci A., Derdiyok A.: Implementation of an adaptive fuzzy compensator for coupled tank liquid level control system. Measurement 91, 12–18 (2016).
- [6] Biswas P., Srivastava R., Ray S., Samanta A.: Sliding mode control of quadruple tank process. Mechatronics 19(4), 548–561 (2009).
- [7] Shah D.H., Patel D.M.: Design of sliding mode control for quadruple-tank MIMO process with time delay compensation. Journal of Process Control 76, 46–61 (2019).
- [8] Truong N.L.V., Lee M.: An extended method of simplified decoupling for multivariable processes with multiple time delays. Journal of Chem. Eng. of Japan 46, 279–293 (2013).
- [9] Chuong, V.L., Vu, T.N.L., Truong N.T.N., Jung J.H.: An analytical design of simplified decoupling Smith predictors for multivariable processes. Appl. Sci. 9, 2487 (2019).
- [10] Bouyedda, H., Ladaci S., Sedraoui, M., Lashab M.: Identification and control design for a class of non-minimum phase dead-time systems based on fractional-order Smith predictor and Genetic Algorithm technique. Int. J. Dynam. Control 7, 914–925 (2019).
- [11] Oustaloup A., Levron F., Mathieu B., et al.: Frequency-band complex non integer differentiator: characterization and synthesis. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. 47(1), 25–39 (2000).
- [12] Igor Podlubny: Fractional-Order Systems and PI<sup>λ</sup>D<sup>μ</sup>-Controllers. IEEE Transactions on Automatic Control 44(1), 208–214 (1999).
- [13] Monje C. A., Chen Y. Q., Vinagre B. M., Xue D.Y., Feliu V.: Fractional-order Systems and Controls. Fundamentals and Applications, Springer-Verlag, London (2010).
- [14] Luo Y., Chen Y.Q., Wang C.Y., Pi Y.G.: Tuning fractional order proportional integral controllers for fractional order systems. J. Process Control 20, 823–831 (2010).
- [15] Beschi M., Padula F., Visioli A.: Fractional robust PID control of a solar furnace. Control Engineering Practice 60, 190–199 (2016).
- [16] De Keyser R., et al.: A novel auto-tuning method for fractional order PI/PD controllers. ISA Transactions 62, 268–275 (2016).
- [17] Vu T.N.L., Lee M.: Analytical design of fractional-order proportional-integral controllers for time-delay processes. ISA Transactions 52, 583–591 (2013).
- [18] Dazi Li, Lang Liu, Jin Q., Hirasawa K.: Maximum sensitivity based fractional IMC-PID controller design for non-integer order system with time delay. J. Process Control 31, 17–29 (2015).
- [19] Amoura K, et al.: Closed-loop step response for tuning PID-fractional-order-filter controllers. ISA Transactions 64, 247–257 (2016).
- [20] Li M., Zhou P., Zhao Z., Zhang J.: Two-degree-of-freedom fractional order-PID controllers design for fractional order processes with dead-time. ISA Transactions 61, 147–154 (2016).
- [21] Lennart Ljung: System Identification: Theory for the User, 2<sup>th</sup> Edition, Prentice Hall, 1999.
- [22] Paul M.J., Van den Hof, Bombois X.: System Identification for Control, Lecture Notes DISC Course, March 2004.
- [23] Y. Zhu: Multivariable System Identification for Process Control, Elsevier Science & Technology Books, 2001.
- [24] Garrido, J., Vázquez, F., Morilla, F.: Inverted decoupling internal model control for square stable multivariable time delay systems. J. Process Control 24, 1710–1719 (2014).
- [25] Skogestad S., Postlethwaithe I.: Multivariable feedback control analysis and design. 1st Edition, John Wiley & Sons, 1996.



# Treatment of Wastewater Containing Reactive Dyes by electro-Fenton Method

Nguyen Thai Anh  
Faculty of Chemical and Food  
Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
anhnt@hcmute.edu.vn

Tran Tien Khoi  
Department of Environmental  
Engineering  
International University, Vietnam  
National University Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
tkhoi@hcmiu.edu.vn

Nguyen Nhat Huy  
Faculty of Natural Resources and  
Environment  
Ho Chi Minh City University of  
Technology, Vietnam National University  
Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
nnhuy@hcmut.edu.vn

Hoang Thi Ngoc Mai  
Faculty of Chemical and Food Technology  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
htngocmai1310@gmail.com

Nguyen Hong Ngoc Linh  
Faculty of Chemical and Food Technology  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
ngoclinhnguyenhong@gmail.com

**Abstract**— The textile dyeing wastewater pollutes the environment with the presence of residual highly reactive dyes. In this study, real textile dyeing wastewater collected from Textile and Dyeing Kim Thanh Hung (KTH) Company and synthetic wastewater prepared from Suncion Red (SR) and Suncion Blue (SB) reactive dyes were used. Fourier-transform infrared spectroscopy analytical results showed that electro-Fenton was effective in textile dyeing wastewater treatment due to its ability to decompose and transform dyes and cyclic compounds. From experiments on effect of  $\text{FeSO}_4$  and  $\text{NaCl}$  amounts, it is discovered that the removal efficiencies of color and COD fluctuated depending on the amounts of two chemicals. And, the suitable conditions for high color removal in batch electro-Fenton model including a 24V-10A of DC power supply and a 100  $\text{cm}^2$  of MMO electrode was found to be in pH range 2 – 4, with the dosage of  $\text{FeSO}_4$  and  $\text{NaCl}$  are 1 and 15 g, respectively for 3 L of wastewater. Besides, the efficiencies of color and COD removal from SR dyes were higher than those from SB dyes. On the contrary, the SB's decolorization rate was faster than that of SR through the difference of the rate constant obtained from kinetics study.

**Keywords**—textile wastewater treatment, electro-Fenton, reactive dyes, advanced oxidation process

## I. INTRODUCTION

The modern trend towards globalization makes it all the more necessary for countries around the world, including Vietnam, to develop socio-economic background. Accordingly, Vietnam's economy is gradually developing, especially the development of industries including textile and garment industry which has exerted adverse effects on the environment due to its wastewater with high concentrations of pollutants. According to results in recent study by Sahunin, et al. [1], approximately 21 – 377  $\text{m}^3$  of water is needed to manufacture one ton of textile fibers, thereby generating a large amount of wastewater, which was mainly from dyeing process and finishing products [2]. The high content of chemical oxygen demand (COD), total dissolved solids (TDS), and residual free chlorine (RFC) are typical characteristics of textile dyeing wastewater [3]. Temperature

of textile dyeing wastewater is from 35 – 45 °C [3], which is a factor affecting treatment efficiency. Therefore, textile dyeing wastewater needs to be cooled before entering the treatment system. In addition, color is one of the serious environmental problems caused by textile dyeing wastewater, which can be up to 2,500 Pt-Co [3]. Dyes used in production process are mostly lost and discharged to the wastewater flow. This is the reason for fluctuating pH (6 – 10) [3] which depends on the kind of dyes. According to Pagga and Brow, 1986 [4], 53% out of 87 dyes are nonbiodegradable. On the other hand, the study on toxicity of 3,000 types of dye showed that 37% were toxic and 2% were very toxic and extremely toxic to aquatic species [5].

Electro-Fenton has been reported to have potential for textile dyeing wastewater treatment [6]. Some values, such as pH, iron(II) ion ( $\text{Fe}^{2+}$ ) concentration (mM), and current density ( $\text{mA}/\text{cm}^2$ ), are the main factors affecting electro-Fenton process. Nidheesh and Gandhimathi, 2012 [7] found that, the optimum pH for electro-Fenton process in wastewater treatment varied from 2 to 4, in which pH 3 was needed for textile dyeing wastewater treatment [8, 9]. Depending on the characteristics of the wastewater, the optimum value of  $\text{Fe}^{2+}$  concentration and current density can be changed [7]. In a study by Cruz-González et al., 2012 [10], the efficiency of dye decolorization reached 95.9% at the optimal operating conditions (i.e., 0.3 mM  $\text{Fe}^{2+}$  and 15  $\text{mA}/\text{cm}^2$ ).

This study aimed to determine the best values of pH and chemical concentration for textile dyeing wastewater treatment by electro-Fenton. The results obtained in this study were basics to develop a technology for textile dyeing wastewater treatment in Vietnam.

## II. EXPERIMENTAL

The chemicals, including  $\text{NaCl}$ ,  $\text{FeSO}_4$ ,  $\text{H}_2\text{SO}_4$ ,  $\text{NaOH}$ ,  $\text{K}_2\text{Cr}_2\text{O}_7$ , and feroin, used in this study were provided by Merck with 99% purity. Real wastewater was collected from Textile and Dyeing Kinh Thanh Hung Company located in

Xuyen A Industrial park. Commercial Suncion Red (SR) and Suncion Blue (SB) reactive dyes were used to simulate textile dyeing wastewater (synthetic wastewater). The synthetic wastewater was prepared according to the following steps: add 1.08 g dyes (SR or SB) to 900 mL of water, stir the mixture, adjust pH to 11 to dissolve dyes completely, follow boiling for 2 h at 100 °C, allow the samples to be cooled at room temperature, and finally dilute with distilled water to obtain a dye concentration of 200 mg/L. Electro-Fenton model (**Figure 1**) was set up with MMO electrode of 100 cm<sup>2</sup> in area. The power used was direct current (10 A, 24 V). The effects of operational condition were investigated with different solution pH values (2 – 8), FeSO<sub>4</sub>:NaCl ratios (1:3, 1:4, 1:5, 1:6, 1:7, 1:8, 1:9, and 1:10), and NaCl amounts (4000 – 6000 mg/L). In case of real wastewater, the pretreatment by coagulation was performed to reduce the level of pollution as well as create the suitable conditions for the inlet of electro-Fenton system. This pretreatment work was also optimized with different pH values (5 - 8) and FeSO<sub>4</sub> concentration (500 – 700 mg/L) in Jartest model.



Fig. 1. Experimental set-up for electro-Fenton treatment of dye wastewater

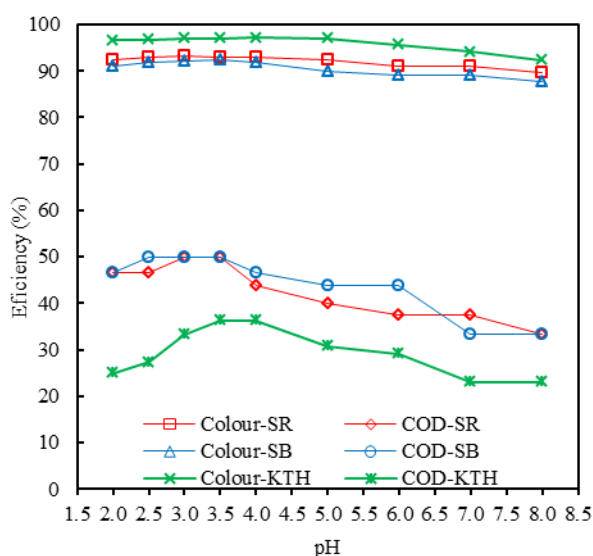


Fig. 2. Changes in color and COD removal efficiencies according to pH of pseudo-wastewater (SR and SB) and real wastewater (from Textile and Dyeing Kim Thanh Hung Company (KTH))

### III. RESULTS AND DISCUSSION

#### A. Effect of Solution pH

pH is an important parameter affecting the electro-Fenton. In this experiment, synthetic wastewater (prepared from SR or SB) with the initial dye concentration of 200 mg/L and COD of 960 to 1024 mg/L, and real wastewater from Textile and Dyeing Kim Thanh Hung (KTH) Company with COD of 768 to 832 mg/L were used. **Figure 2** illustrates how pH affects removal efficiencies of color and COD in synthetic and real wastewaters. As the figure shows, removal efficiencies of color were higher than those of COD that did not depend on solution pH. Efficiencies achieved at low pH were better than those achieved at high pH and the best efficiency reached in the pH range of 2 – 4. Because the H<sup>+</sup> concentration at low pH (< 2) was high, meaning the stronger competition of the H<sup>+</sup> reduction to H<sub>2</sub>, the efficiency of the redox process producing H<sub>2</sub>O<sub>2</sub> was low. At high pH (> 4), the H<sup>+</sup> concentration was not enough to form H<sub>2</sub>O<sub>2</sub>. Fe<sup>2+</sup> concentration in solution decreased because of the conversion from Fe<sup>2+</sup> to Fe<sup>3+</sup>. The results of this experiment were found to be consistent with those of Nidheesh and Gandhimathi, 2012 [7]. Regarding the optimum pH for removal of color and COD, it is obvious that the best removal efficiencies of color and COD in the sample prepared from SR were 93.26% and 50%, respectively at pH 3. Meanwhile, those of color and COD in the sample prepared from SB were 92.30% and 50%, respectively at pH 3.5. With real wastewater which was pretreated by coagulation, in the electro-Fenton step, the optimum value of pH was 4, at which color and COD removal efficiencies reached the highest values of up to 97.17% and 36.36%, respectively.

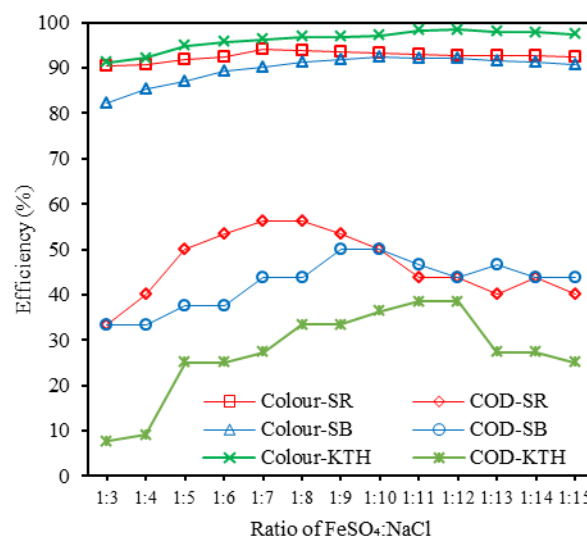


Fig. 3. Changes in color and COD removal efficiencies according to FeSO<sub>4</sub> amount

#### B. Effect of FeSO<sub>4</sub> Amount

In this experiment, concentration of NaCl was remained constant at 5 g/L for synthetic wastewater and at 4 g/L for real wastewater. Thus, changes in the amount of FeSO<sub>4</sub> would result in changes in FeSO<sub>4</sub>:NaCl ratio. From the results in **Figure 3**, it is apparent that the removal efficiencies of color-SR and COD-SR were higher than those of color-SB and COD-SB, respectively. The best removal efficiencies of COD-SR and COD-SB reached at the

FeSO<sub>4</sub>:NaCl ratio of 1:7 and 1:10 at the best pH in previous tests, respectively. This ratio for real wastewater, on the other hand, was 1:12, at which color and COD removal efficiencies reached a peak of 98.44% and 38.46%, respectively.

In all three cases, treatment efficiency tended to decrease as FeSO<sub>4</sub> amount used increased. The reason was that the Fe<sup>2+</sup> concentration was too small, not enough to react with the H<sub>2</sub>O<sub>2</sub> generated at the cathode, so OH<sup>•</sup> concentration was low. When the amount of Fe<sup>2+</sup> was high, the oxidation of Fe<sup>2+</sup> at the anode to create Fe<sup>3+</sup> occurred, leaving an excess of Fe<sup>3+</sup>/Fe<sup>2+</sup> redox couple, thereby making redox cycle on both electrodes happen continuously and reducing the removal efficiency.

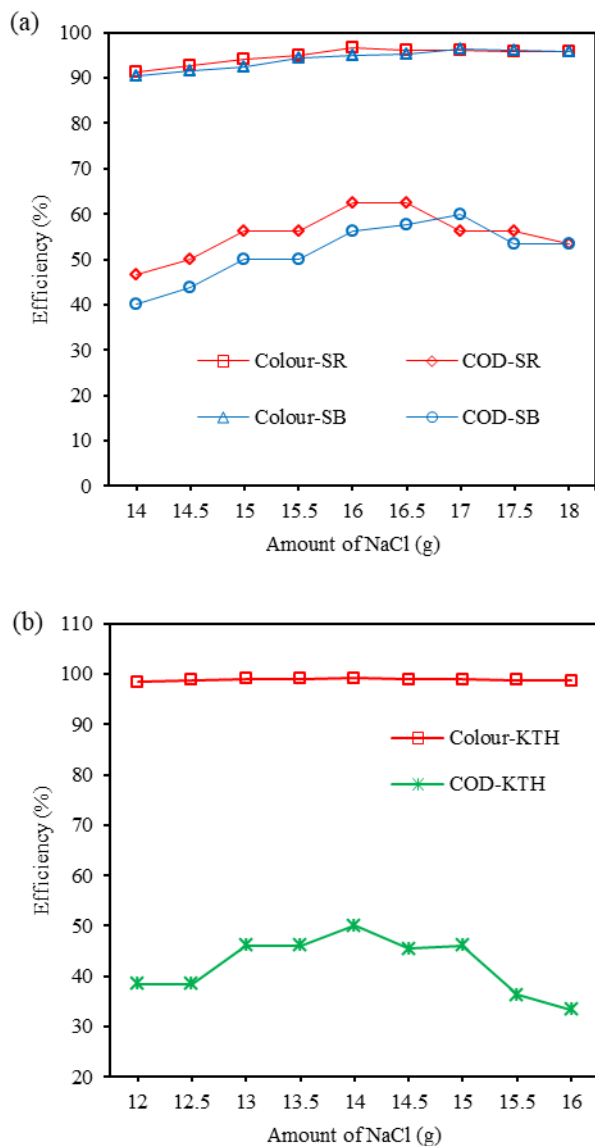


Fig. 4. Changes in color and COD removal efficiencies according to NaCl amount in (a) pseudo-wastewater and (b) real wastewater

### C. Effect of NaCl Concentration

It is known that the amount of NaCl used directly affects the processing efficiency. When the amount of NaCl increases to the optimum point, the amount of H<sub>2</sub>O<sub>2</sub> produced is sufficient and the Fe<sup>3+</sup>/Fe<sup>2+</sup> redox couple is sufficient, limiting the effect of OH<sup>•</sup> creation. If NaCl is less than the optimal amount, it will affect the current in the

solution, causing instability in the process. In contrast, if NaCl is too much compared to the optimal amount, the process will continue, but at this time the process is no longer electro-Fenton and becomes electrochemical flocculation.

The results in **Figure 4** show that the best amount of NaCl was different to each experiment with SR wastewater, SB wastewater, and real wastewater. In particular, in experiment with SR wastewater (**Figure 4a**), the best NaCl amount was 16 g (at pH 3 and 2.14 g FeSO<sub>4</sub>), color and COD removal efficiencies were 96.7% and 62.5%, respectively. With SB wastewater (**Figure 4a**), the best NaCl amount was 17 g (at pH 3.5 and 1.5 g FeSO<sub>4</sub>), color and COD removal efficiencies were 96.51% and 60%, respectively. The 14 g NaCl amount, 99.14% of color, and 50% of COD were the best values reached in the experiment with real wastewater (**Figure 4b**).

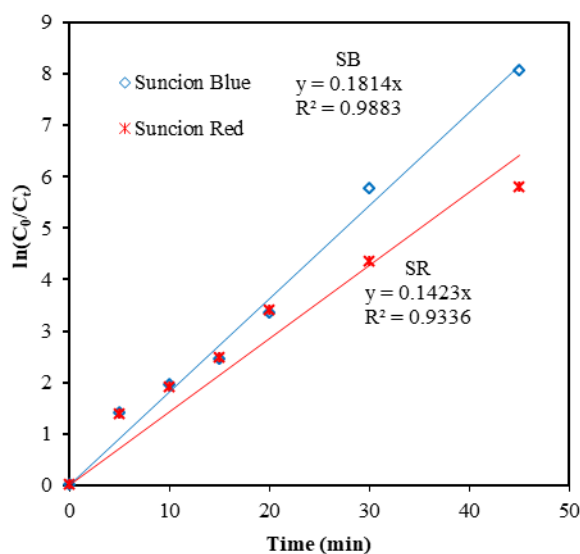


Fig. 5. First-order kinetic graph for SR, SB dyes:  $C_0 \approx 200$  mg/L

### D. Kinetics of the Electro-Fenton Treatment Process

As can be seen from data depicted in **Figure 5**, the first-order equation was quite suitable to describe the kinetics of SR and SB dyes treated by electro-Fenton because the correlation coefficient  $R^2$  was greater than 0.9. Comparing the rate constants in the two equations shows that dyes with high processing efficiency was unnecessarily decomposed faster. Furthermore, **Figure 2** and **Figure 3** displayed that the treatment efficiency of color is much higher than that of COD. The big differences in color removal, COD reduction and rate constant implied that the color-making components decomposed faster than the remaining components that made the COD.

**Figure 6** shows the results of FTIR in this study. Various peaks were identified at wavelength numbers of  $< 900$  cm<sup>-1</sup> (ring compound), 1047 (-S=O), 1208 (aromatic), 1542 (diazo), 1578 (-N=N- stretching of azo group), 2924 (symmetric stretching), 2853 (asymmetric stretching) and 3422 (N-H of dye structure) in the structure of dye before treatment [11, 12]. However, these peaks were all disappeared after treatment by electro-Fenton, indicating that the dyes and cyclic radicals were decomposed and transformed into new groups with simpler structures.

### E. Role of coagulation

After conducting experiment on real wastewater by coagulation with iron salt  $\text{FeSO}_4$ , the results show that the experiment achieved the best efficiency at pH 7.5 and 5.5 mL of  $\text{FeSO}_4$  for 400 mL wastewater. After 68.09% of color was removed by coagulation, wastewater was continued to be treated by electro-Fenton.

As a result, the color removal efficiency by electro-Fenton combined coagulation was higher than by single electro-Fenton (99.69% > 90.26%) after 5 min. It is transparent that after 5 min, there was no big difference in efficiencies between the two methods. Looking at the entire process, electro-Fenton method after 30 min could completely treat the color while it only took about 10 min for the combined method to completely treat the color. Therefore, each method had its own advantages and disadvantages that were different from the other. The electrochemical method costed more electricity than the combined method while the combined method took more time and chemicals.

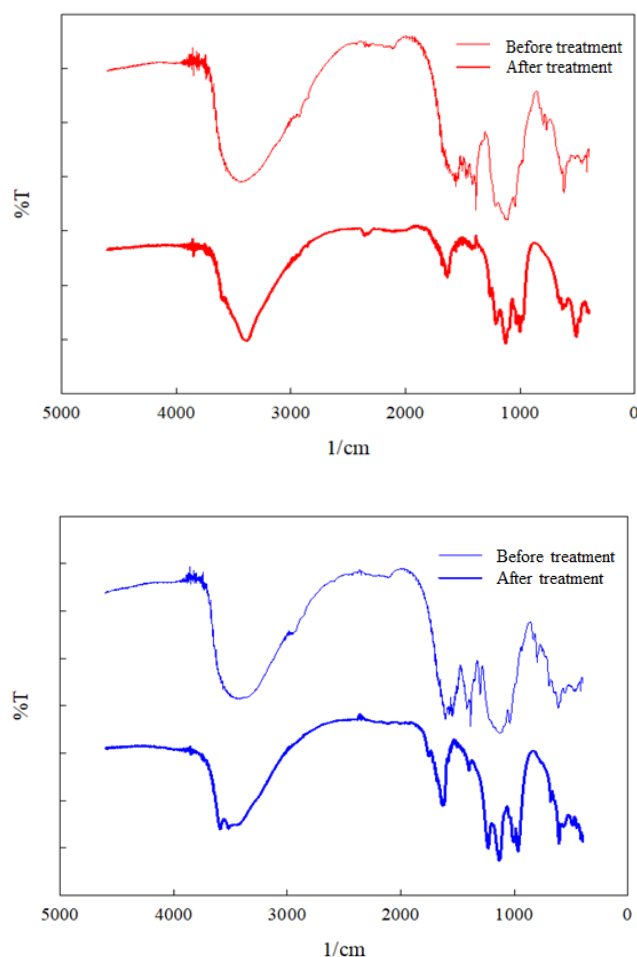


Fig. 6. FTIR spectroscopy of (a) SR and (b) SB dyes before and after treatment by electro-Fenton

### IV. CONCLUSIONS

Electro-Fenton method had the ability to remove over 90% of color in textile dyeing wastewater and synthetic

wastewater prepared from high reactive dyes (Suncion Red or Suncion Blue). pH significantly affected the color and COD removal efficiencies. The efficiencies peaked at pH 2 – 4 and decreased as pH increased. The combined method (electro-Fenton and coagulation) had the color removal efficiency higher than the electro-Fenton. However, it depended on the treatment requirements and actual conditions to consider which treatment method was selected to use. More detailed research is needed in narrower selective intervals so that the error is lowest and the optimal conditions for the electro-Fenton process can be evaluated more accurately.

### ACKNOWLEDGMENT

This work belongs to the project grant No: T81-A/GCN-KHCN. funded by Ho Chi Minh City University of Technology and Education, Vietnam.

### REFERENCES

- [1] C. Sahunin, J. Kaewboran, and M. Hunsom, "Treatment of Textile Dyeing Wastewater by Photo Oxidation using UV/H<sub>2</sub>O<sub>2</sub>," *reactions*, vol. 22, p. 23, 2006.
- [2] N. Rosli, "Development of biological treatment system for reduction of COD from textile wastewater," *Malaysia: Universiti Teknologi*, 2006.
- [3] S. S. Kalra, S. Mohan, A. Sinha, and G. Singh, "Advanced oxidation processes for treatment of textile and dye wastewater: a review," in *2nd International conference on environmental science and development*, 2011, pp. 271-5.
- [4] U. Pagga and D. Brown, "The degradation of dyestuffs: Part II Behaviour of dyestuffs in aerobic biodegradation tests," *Chemosphere*, vol. 15, pp. 479-491, 1986.
- [5] M. R. Haris and K. Sathasivam, "The removal of methyl red from aqueous solutions using banana pseudostem fibers," *American Journal of applied sciences*, vol. 6, p. 1690, 2009.
- [6] L. Labiadh, M. A. Oturan, M. Panizza, N. B. Hamadi, and S. Ammar, "Complete removal of AHPs synthetic dye from water using new electro-fenton oxidation catalyzed by natural pyrite as heterogeneous catalyst," *Journal of hazardous materials*, vol. 297, pp. 34-41, 2015.
- [7] P. Nidheesh and R. Gandhimathi, "Trends in electro-Fenton process for water and wastewater treatment: an overview," *Desalination*, vol. 299, pp. 1-15, 2012.
- [8] C.-T. Wang, J.-L. Hu, W.-L. Chou, and Y.-M. Kuo, "Removal of color from real dyeing wastewater by Electro-Fenton technology using a three-dimensional graphite cathode," *Journal of hazardous materials*, vol. 152, pp. 601-606, 2008.
- [9] C.-T. Wang, W.-L. Chou, M.-H. Chung, and Y.-M. Kuo, "COD removal from real dyeing wastewater by electro-Fenton technology using an activated carbon fiber cathode," *Desalination*, vol. 253, pp. 129-134, 2010.
- [10] K. Cruz-González, O. Torres-Lopez, A. M. García-León, E. Brillas, A. Hernández-Ramírez, and J. M. Peralta-Hernández, "Optimization of electro-Fenton/BDD process for decolorization of a model azo dye wastewater by means of response surface methodology," *Desalination*, vol. 286, pp. 63-68, 2012.
- [11] D. Kalyani, A. Telke, R. Dhanve, and J. Jadhav, "Ecofriendly biodegradation and detoxification of Reactive Red 2 textile dye by newly isolated *Pseudomonas* sp. SUK1," *Journal of Hazardous Materials*, vol. 163, pp. 735-742, 2009.
- [12] R. Dhanve, U. Shedbalkar, and J. Jadhav, "Biodegradation of diazo reactive dye Navy Blue HE2R (Reactive Blue 172) by an isolated *Exiguobacterium* sp. RD3," *Biotechnology and Bioprocess Engineering*, vol. 13, pp. 53-60, 2008.

# Breast Image Segmentation for evaluation of Cancer Disease

Thanh-Tam Nguyen  
Faculty of Biomedical Engineering  
International University, VNU,  
Ho Chi Minh City, Vietnam  
nttam@hcmiu.edu.vn

Thanh-Hai Nguyen  
Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education,  
Ho Chi Minh City, VietNam  
nthai@hcmute.edu.vn

Ba-Viet Ngo  
Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education,  
Ho Chi Minh City, Vietnam  
vietnb@hcmute.edu.vn

Duc-Dung Vo  
Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education,  
Ho Chi Minh City, VietNam  
dungvd@hcmute.edu.vn

**Abstract**—Breast cancer is one of dangerous diseases and difficult to cure. It is observed that early detection of malignancy can help in the diagnosis of the disease and patient can be saved. For the detection of breast cancer, breast images will be enhanced using a Fuzzy logic and possibility distribution algorithm and then segmented to produce images with region of interest, in which just cancer shape appears in the image for detecting and estimating disease status. This paper proposes a statistic method based on the gray level of pixels in the image through histograms of two breast image sets to classify two cases of cancer and normal one. Simulation results on breast image sets will show that the proposed method is effective and it can be developed for detection of benign and malignant tumors in artificial intelligent systems.

**Keywords**— Breast image sets; Image enhancement; Image segmentation; Histogram for evaluation; Statistics of contrast

## I. INTRODUCTION

Cancer is one of the most dangerous and very difficult diseases to treat for one patient. In types of cancer, Breast Cancer (BC) often appears in women, the most women are 40 years old. In current, there are about 164,671 breast cancer diseases in Vietnam and about 114,871 people were died [1]. In addition, BC has the third high rate (9.2%) and is behind liver and lung cancers. In particular, BC in women accounted about 15,229 of the 73,849 cancer cases according to Globalcan statistics in 2018. Therefore, BC is possibly diagnosed early, the chance of complete cure is very high. Thus, using methods of image processing for analysing and diagnosing diseases early are very important.

Diagnostic imaging doctors often use various imaging methods including Computerized Tomography (CT) [2], ultrasound [3], mammography [4] and Magnetic Resonance Imaging (MRI) [5] for screening and diagnosis to detect cancer disease early. In these methods, ultrasound and x-ray images mainly used to detect and diagnose breast cancer. The blurred image edges and the low contrast of the ultrasound image are one challenge in automatic image

segmentation. During capturing x-ray breast image, high-resolution images with low-energy X-rays allow to detect abnormalities or tumors obscured or overlapped by surrounding breast tissue [6-7]. To extract anomalies or areas of interest from x-ray breast images, image segmentation can be applied. In practice, there are various segmentation and in diagnostic techniques, pre-processing needs to be applied to remove labels, tags, patient names or other unwanted information. In addition, these techniques increase image contrast and eliminate noise for making the segmented images more accurate and reliable.

After image preprocessing, image segmentation algorithms play an important role in determining whether tumor in image is malignance or benign. To detect malignant tumor, features such as intensity, shape, size, texture, gray scale histogram for describing the tumor can be calculated. These segmentation algorithms can be classified into groups including regional approach (grouping pixels into homogeneous regions during large computation based on high resolution) [8]; based on contours (meaning based on discontinuity of color, gray level or texture of image edges detected respectively), based on cluster (pixel clusters with the same property) [9]; threshold method (foreground segmentation from background through information from gray level histogram); methods based on energy function [10].

Region-based algorithm applied on mini-MIAS database for chest muscle segmentation was the 98% accuracy [11-12] and in the case of EPIC data sets, this algorithm had the 91.5% accuracy. The accuracy of this algorithm in removing noise and extraneous components from the mini-MIAS dataset was very good, about 98.8% [13]. However, this technique is mostly used due to its high resolution, so it takes more time during segmentation as well as requiring the selection of same points is difficult. Another technique is based on analyzing the energy of components in the image to eliminate unnecessary components for enhancing the image and removing chest muscle from image, with the 90.37%



accuracy [14]. This technique is very flexible and just requires little calculation.

Clustering techniques were proposed to apply in methods of SVM, fuzzy c-means and decision trees [15-21], in which they could work well for overlapping data, giving high accuracy in clustering and detecting breast tumors from images having sensitive problems to initial clusters and peripheral values. In addition, clustering and texture filters can effectively detect calcification even with small noise image or large tumors inside the breast. Moreover, the c-mean fuzzy clustering technique has large tolerance for contrast, fuzzy boundaries, noise and this can produce high precision in image segmentation. Therefore, this method is better for segmenting breast tumor lesions.

This paper is organized as follows. Section I is introduction about methods of enhancement, segmentation, binary conversion related to classification of breast images. In Section II, the paper will present methods of enhancement, segmentation for searching ROI, binary conversion and gray level statistics for searching features of breast images. The objective of Section III is that simulation results and discussion are expressed, in which a statistic of breast image sets is produced to evaluate cancer images. Finally, conclusion will be shown in this Section IV.

## II. MATERIAL AND METHOD

### A. Algorithm for Image Enhancement

Breast cancer disease is very dangerous and very difficult to detect it in the first stage. This difficulty can be due to two reasons: patient is not regularly screened to detect cancer and breast image sometimes does obvious for diagnosis. Therefore, processing breast images to detect breast cancer soon is very necessary. In this paper, the source of datasets was collected from Mammographic Image Analysis Society (MIAS). Original images are often hard to accurately diagnose breast image status with cancer or normal. Therefore, the images can be enhanced before segmentation for detecting Region of Interest (ROI) and then statistics for evaluation of disease status as shown in Fig. 1.

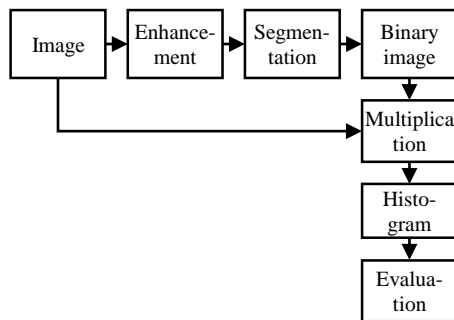


Fig. 1. Block diagram of cancer evaluation process

After collecting the datasets from MIAS, the input images are enhanced using a Fuzzy logic and possibility distribution algorithm [21]. In the research, the minimum, maximum and mean values of pixels of one gray level image are calculated and then two thresholds of  $Th_1$ ,  $Th_2$  are calculated as follows:

$$Th_1 = (mean + min)/2 \quad (1)$$

$$Th_2 = (mean + max)/2 \quad (2)$$

From the two threshold values, the gray level image is divided into four groups for determining four corresponding thresholds as follows:

$$min \leq P_0 < Th_1$$

$$Th_1 \leq P_1 < mean$$

$$mean \leq P_2 < Th_2$$

$$Th_2 \leq P_3 \leq max$$

in which,

- $P_0, P_1, P_2, P_3$  are the new gray level thresholds
- $min$  denotes the smallest gray level value in the image
- $max$  is the largest gray level value
- $mean$  describes the gray level average value

The gray level of pixels in each group is adjusted and then all pixels in the image are calculated to correspond to gray levels using the following equations:

$$P_{N0} = 2 \times \left( \frac{(P_0 - min)}{(mean - min)} \right)^2 \quad (3)$$

$$P_{N1} = 1 - 2 \times \left( \frac{(P_1 - mean)}{(mean - min)} \right)^2 \quad (4)$$

$$P_{N2} = 1 - 2 \times \left( \frac{(P_2 - mean)}{(max - min)} \right)^2 \quad (5)$$

$$P_{N3} = 2 \times \left( \frac{(P_3 - mean)}{(max - mean)} \right)^2 \quad (6)$$

where  $P_{N0}, P_{N1}, P_{N2}, P_{N3}$  are four new thresholds calculated based on previous thresholds and the values of  $min$ ,  $max$  and  $mean$  in the breast image.

From Eq. (3) to Eq. (6), all breast images are enhanced and contrast levels are adjusted. It is obvious that the breast images after enhancement for image segmentation are better to search ROI areas.

### B. Otsu Image Segmentation

After enhancement of images, an Otsu segmentation algorithm is applied to determine thresholds of gray level for convert gray level images to binary images [22]. In particular, the Otsu segmentation is to determine gray level, where a gray level image is calculated to divide pixels into two groups: background pixels, and object pixels. Moreover, the threshold is calculated to minimize the intra-class variance of two classes (background and object). In order to search this threshold, the algorithm is described as follows:

1. Calculate histogram and probabilities of each class in image
2. Set up initial probabilities
3. Update probabilities and calculate mean values of two classes. Step through all possible
4. Desired threshold corresponds to the maximum value of

The Otsu threshold is applied to convert the gray level image into a binary image, in which the ROI area is 1s pixels and the background is 0s pixels. For representation of ROI histogram, the binary image is multiplied to the original image to produce the gray level ROI. Therefore, the probability density of pixels in the image with the gray ROI will produce a gray level histogram, which allows us calculate the different gray level area between two types of breast image (cancer and normal). From these different gray level areas of the cancer and normal images, we can calculate to determine breast cancer image.

### C. Determination of Breast Image

In this paper, statistics of contrast and corresponding pixel amount in two breast image sets are performed based on gray level histograms of two types of breast image (cancer and normal). These statistic represents two types of gray level corresponding to two groups of cancer and normal breast images which can be considered as features of these types. Therefore, we can determine an ability of breast cancer disease based these features. The contrast  $C$  of gray level ROI and the corresponding pixels in a breast image sets are determined by using the following formulas:

$$C = \frac{1}{NM} \sum_{i=1}^M \sum_{j=1}^N [(i-j)^2 p(i,j)] \quad (7)$$

in which  $N$ ,  $M$  are the dimensions of the Gray-Level Co-Occurrence Matrix (GLCM) of the image, and  $p(i,j)$  is the frequency related to the gray levels  $i$ ,  $j$  of two adjacent pixels.

## III. RESULTS AND DISCUSSION

In this research, simulation results are worked out from original breast image sets using algorithms of enhancement, segmentation, binary image conversion and statistics based on gray level histograms, in which the breast image sets consist of ten normal images and ten cancer ones.

### A. Results of Image Enhancement

For enhancement of a breast image, *min*, *max* and *mean* values in each image are calculated for determining thresholds of  $P_{N0}$ ,  $P_{N1}$ ,  $P_{N2}$ ,  $P_{N3}$ . Fig.2 shows original and enhanced breast images. It is obvious that the breast images after enhancing show gray level areas which we can identify cancer status through next processing.

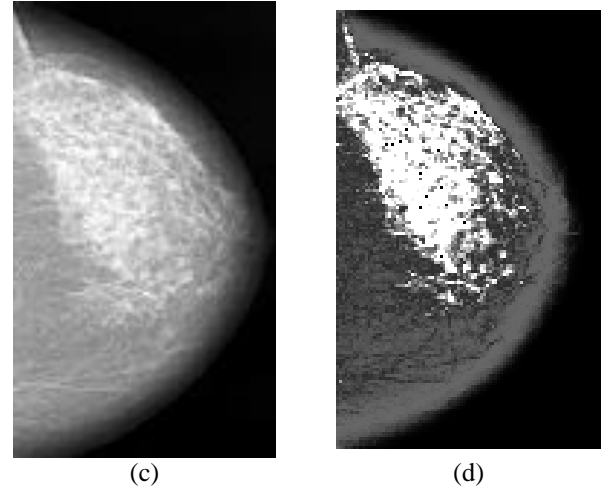
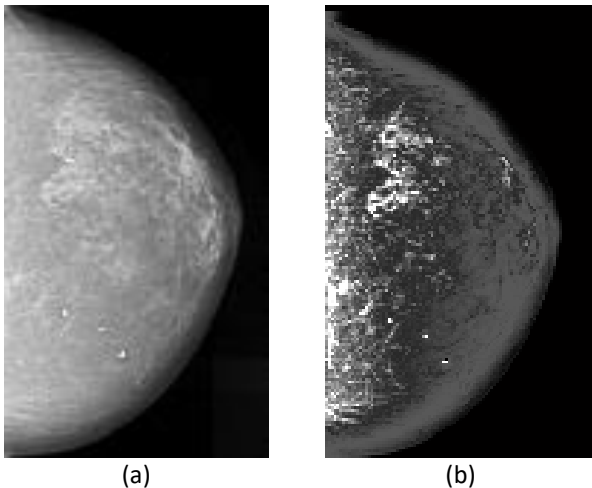


Fig. 2. Representation of original and enhanced images

- (a) original image of normal case
- (b) enhanced image of normal case
- (c) original image of cancer case
- (d) enhanced image of cancer case

### B. Results of Image Segmentation

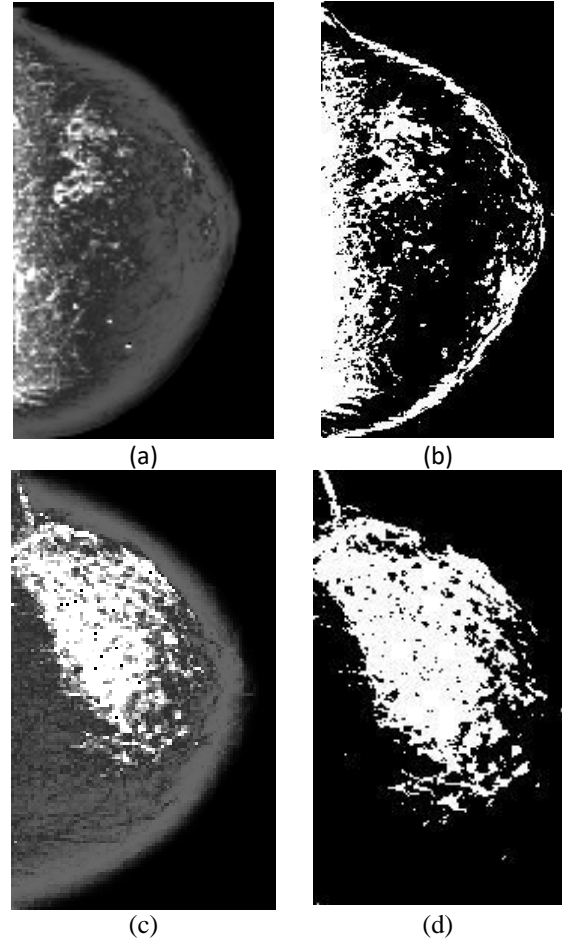


Fig. 3. Representation of enhanced and segmented images

- (a) enhanced image of normal case
- (b) segmented image of normal case
- (c) enhanced image of cancer case
- (d) segmented image of cancer case

From enhanced breast images, the Otsu segmentation method was applied to produce threshold which allow to separate ROI for evaluation of image status. Breast images after segmentation were converted into binary images as shown in Fig. 2. With these binary images, basically we can

see the difference between white areas (0 gray level) and back ones (1 gray level) in two cases of cancer and normal.

### C. Breast Image Evaluation

For evaluation of cancer status, binary images were multiplied with original breast images to produce images with gray level ROI which allow to identify between cancer and normal images. Therefore, probabilities of gray level pixels were calculated to create a gray level histogram for breast image evaluation.

From breast images with the ROI, all breast images were calculated to create gray level histograms. Therefore, the gray level densities between two types breast images (cancer and normal) were different as shown in Fig. 4 and Fig.5. In particular, the pixel density of cancer images has the sudden change in the gray level range of around 0.18-0.20, while that of normal images slowly changes in the same range.

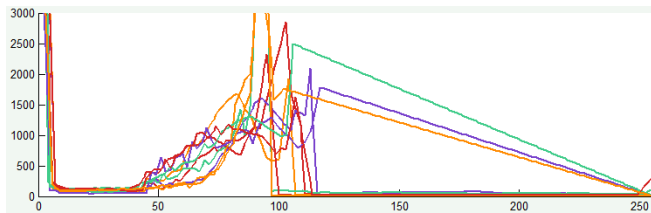


Fig. 4. Gray level histograms of normal case

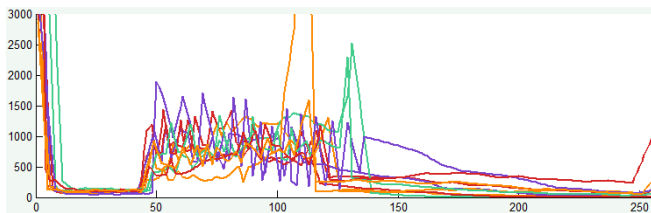


Fig. 5. Gray level histogram of cancer case

Fig. 4 and Fig. 5 show that in cancer cases, pixels corresponding to the gray level range of around 50 to 130 in the histograms change suddenly from low to high, while those corresponding to the same gray level range in normal cases increase slowly. Therefore, we can basically evaluate the difference between two types of breast images. However, contrast values and corresponding pixel amount provided more information for evaluation of breast disease status.

TABLE I. FEATURER EXTRACTION OF CANCER AND NORMAL CASES

Case	Cancer		Normal	
	Contrast	ROI pixels	Contrast	ROI pixels
1	0.1927	558	0.0758	319
2	0.1478	417	0.0711	331
3	0.1582	484	0.0702	207
4	0.1419	402	0.0269	343
5	0.1501	601	0.0458	391
6	0.0817	513	0.0181	212
7	0.2799	527	0.0584	388
8	0.2483	560	0.0855	326
9	0.1185	442	0.0746	98
10	0.1455	420	0.0295	77

Table 1 shows the contrast values corresponding to the average pixel amount of ROIs in cancer and normal images in the gray level range of 0.2-0.3 in the histograms. It is obvious that the contrast values in the normal images are from 0.0181 to 0.0855, while those in the cancer images are

from 0.0817 to 0.2799. In addition, ROI pixels corresponding to image contrast of ten ROI images for each type were different. Particular, ROI pixels in cancer images were the range of 402 to 601, while the range of ROI pixels in normal images was 77 to 391.

With the experiment of 20 breast images including ten image for each type, one can evaluate the ability of disease status based on image after processing. In particular, with this statistic table, we can evaluate breast image status and this can be one of information for doctor in disease diagnosis.

Mammography images were segmented using fuzzy c-means clustering and then ROIs in these image were calculated for determining normal and abnormal regions [21]. Result of the classification accuracy in this research was around 92%. In our research, image processing techniques were applied, histograms of ROIs in mammography images were represented and then statistics of pixel densities on both normal and abnormal images were performed to be the basic for evaluation. Contrast values and corresponding pixel amount were determined for evaluation of disease status.

## IV. CONCLUSION

In this paper, the set of 20 breast images including 10 image for each type were processed to produce features for evaluation of disease status. Therefore, this paper presented the proposed method consisting of the image enhancement using the Fuzzy logic and possibility distribution algorithm, Otsu segmentation and statistics on ROI image based on histograms for evaluation of breast cancer. In particular, from cancer and normal images, features based on contrast values and the corresponding pixel amount were obtained as shown in Table 1, in which the ranges from the minimum and maximum values between two types of normal and cancer images were really obvious. Simulation results showed the effectiveness of the proposed method and also it is very significant for development of breast image recognition using artificial intelligence with higher performance.

## ACKNOWLEDGMENT

The authors would like to acknowledge the support by the HCMC University of Technology and Education, Vietnam.

## REFERENCES

- [1] "Vietnam Source: Globocan 2018", World Health Organization, 5-2019.
- [2] Chen, Biao, and Ruola Ning "Cone-beam volume CT breast imaging: a Feasibility study." *Medical Physics* 29.5 (2002): 755-770.
- [3] Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical imaging* 37.3 (2013): 420-426.
- [4] Olsen, Ole, and Peter C. Gøtzsche. "Cochrane review on screening for breast cancer with mammography" *The Lancet* 358.9290 (2001): 1340-1342.
- [5] Mann, R. M., Kuhl, C. K., Kinkel, K., & Boetes, C. "Breast MRI: guidelines from the European society of breast imaging" *European Radiology* 18.7 (2008): 1307-1318.
- [6] Johns, Paul C., and Martin J. Yaffe. "X-ray characterization of normal and neoplastic breast tissues" *Physics in Medicine & Biology* 32.6 (1987): 675.

- [7] Mustra, Mario, Mislav Grgic, and Rangaraj M. Rangayyan. "Review of recent advances in the segmentation of the breast boundary and the pectoral muscle in mammograms" *Medical & biological engineering & computing* 54.7 (2016): 1003-1024.
- [8] Szeliski, Richard "Computer vision: algorithms and application" Springer Science & Business Media, 2010.
- [9] Muthukrishnan, R., and Miyilsamy Radha "Edge detection techniques for image segmentation," *International Journal of Computer Science & Information Technology* 3.6 (2011): 259.
- [10] Kas, Michael, Andrew Witkin, and D. Terzopoulos. "Snakes-active contours models." *International Journal of Computer Vision* 1.4 (1987): 321-331.
- [11] Raba, D., Oliver, A., Martí, J., Peracaula, M., & Espunya, J. "Breast segmentation with pectoral muscle suppression on digital mammograms," *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, Berlin, Heidelberg, 2005.
- [12] Chen, Zhili, and Reyer Zwiggelaar. "A combined method for automatic identification of the breast boundary in mammograms." *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on*. IEEE, 2012.
- [13] Zhang, Zhiyong, Joan Lu, and Yau Jim Yip. "Automatic segmentation for breast skin-line." *the IEEE 10th International Conference on Computer and Information Technology*, 2010.
- [14] Lu, X., Dong, M., Ma, Y., & Wang, K. "Automatic Mass Segmentation Method in mammograms based on improved VFC Snake model" *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*. 2015. 201-217.
- [15] Anitha, J., and J. Dinesh Peter. "Mass segmentation in mammograms using a kernel-based fuzzy level set method." *International Journal of Biomedical Engineering and Technology* 19.2 (2015): 133-153.
- [16] Touil, Asma, and Karim Kalti. "Iterative fuzzy segmentation for an accurate delimitation of the breast region." *Computer methods and programs in biomedicine* 132 (2016): 137-147
- [17] Feng, Y., Dong, F., Xia, X., Hu, C. H., Fan, Q., Hu, Y., ... & Mutic, S. "An adaptive fuzzy C-means method utilizing neighboring information for breast tumor segmentation in ultrasound images." *Medical Physics* 44.7 (2017): 3752-3760.
- [18] Shi, P., Zhong, J., Rampun, A., & Wang, H. "A hierarchical pipeline for breast boundary segmentation and calcification detection in mammograms." *Computers in biology and medicine* 96 (2018): 178-188.
- [19] Valdés-Santiago, D., Quintana-Martínez, R., León-Mecías, Á., & Díaz-Román, M. L. B. "Mammographic Mass Segmentation Using Fuzzy C-means and Decision Trees." *International Conference on Articulated Motion and Deformable Objects*. Springer, 2018.
- [20] Kashyap, K. L., Bajpai, M. K., Khanna, P., Giakos, G. "Mesh-free based variational level set evolution for breast region segmentation and abnormality detection using mammograms." *International journal for numerical methods in biomedical engineering* 34.1 2018.
- [21] Chowdhary, Chiranjil Lal, and D. P. Acharjya. "Segmentation of Mammograms Using a Novel Intuitionistic Possibilistic Fuzzy C-Mean Clustering Algorithm." *Nature Inspired Computing*. Springer, Singapore, 2018. 75-82.
- [22] Zhang, Jun & Hu, Jinglu (2008). "Image segmentation based on 2D Otsu method with histogram analysis". *Computer Science and Software Engineering*, 2008 International Conference on. 6: 105–108

# Vision-based People Counting for Attendance Monitoring System

Manh Cuong Le

Faculty of Electrical of Electronics and  
Engineering  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
mcuongle1206@gmail.com

My-Ha Le\*

Faculty of Electrical of Electronics and  
Engineering  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
Corresponding author:  
halm@hcmute.edu.vn

Minh-Thien Duong

Faculty of Electrical of Electronics and  
Engineering  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
duongthien2206@gmail.com

**Abstract**—In this paper, an automatic people detection and counting system using data collected from an over-head camera is proposed. The purpose of this research is to develop a fast and accurate intelligent people counting technique for attendance monitoring systems in offices and lecture rooms. The proposed method includes two stages working sequentially. First, the detection task is executed to find any person presented in the current frame. A deep learning architecture, MobileNetv2-SSD, was used to carry out the detecting phase. If there is any detected person, the tracking phase, which based on visual-tracking techniques, will be initialized, and keeps track of the people's position. Based on the tracked motion path of the detected people, we can determine if there is any person who has entered or exited the room. Therefore, we can monitor the number of attending people. The testing hardware was a Raspberry Pi computer and a camera. This work has been tested on different stages of a day and achieved real-time performance with sufficient accuracy.

**Keywords**—people counting, human detection, object tracking

## I. INTRODUCTION

People counting using intelligence camera systems play an important role in many different applications related to security, marketing, and surveillance. Especially, this technique could make a great contribution to the “*Social Distancing*”, which one of the most efficient solutions for the world pandemic in 2020, by aiding the problem of monitoring the number of people in a specific area.

Many people-counting based on computer vision researches have been done in the past few years. [1] used a background subtraction between frames to track moving objects - human in this case, and did the people-counting based on the tracked people. [2], [3] evaluate the number of people move under an over-head camera using head detection and tracking. However, previous approaches have a lot of limitations in the detecting phase such as using image processing techniques and simple classifiers, which are not robust in a variety of illuminating conditions and different environments.

In this paper, a different approach for people-counting which used a combination of detection and tracking techniques is proposed. The system used a camera for acquiring top-down footage of walking people near the entrance of a room. The detection process is done using a Deep Learning architecture, a combination of MobileNetv2 [4] and SSD [5], which achieved state-of-the-art results in

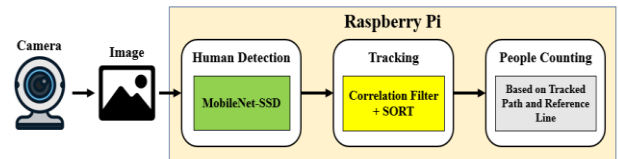


Fig. 1. Block diagram of the automatic people counting system

object detection. Next, a tracking algorithm will associate the next successful detection with the previous one to maintain the identity of a tracked person. However, this tracking method requires a strong detection response in the next frames, a visual tracking technique was integrated to work along with the detection model to prevent the miss-detect situation. A line is created parallel to the entrance door as the reference, and with the tracked path of the detected people, we can determine whether a person has entered or exited the room.

The main advantage of the proposed system is combining the high accuracy Deep-learning-based object detection (human in this case) with a very fast visual tracking method, which can effectively generate position information of people moving near the entrance for the counting algorithm. The workflow diagram of the system is demonstrated in Figure 1.

The rest of the paper proceeds as follows. In Section II, the proposed method is discussed, including MobileNetv2-SSD architecture, the tracking technique based on correlation filter, and the SORT [6] algorithm. The training and application of the proposed method is presented in Section III. Conclusions are drawn in Section IV.

## II. PROPOSED METHODS

The proposed system includes three different algorithms work sequentially. First, the Deep Learning model, named MobileNetv2-SSD, takes the image captured by the camera as the input data, then makes predictions of every human object presented in the frame. For each successful prediction, a visual tracking filter is initialized using the corresponding coordinates of the person. This filter, which based on correlation operation, will keep track of the human object in the next images. The SORT tracking algorithm will associate the identity of all the detected or tracked persons in the frame sequence and based on the tracked IDs, we can determine if people have gone in or out of the room.



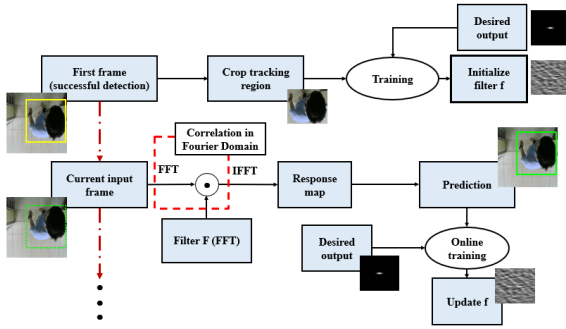


Fig. 2. Correlation Filter based Visual Tracking

### A. Human Detecting and Tracking

In order to monitor the number of people, the system needs to have the ability to detect every person walking near the entrance. Deep-learning based detectors have been proved to be the most robust and reliable method for object detection in general. For the detection task, we fine-tuned a pre-trained MobileNetv2-SSD model [4] on our custom dataset. This detection model includes two convolutional neural networks working along: a feature extractor and a detection framework. We will discuss this further below. Figure 2 illustrates the architecture of the MobileNetv2-SSD model.

#### 1) Feature Extractor (MobileNetv2)

MobileNet-v2 [4] is one of the MobileNet based classification models. The Depthwise Separable Convolution layer, which was introduced in [7], significantly reduce the computational speed but still maintain relatively high classification accuracy, compared to other models such as VGG [8], ResNet [9], etc. The Depthwise Separable Convolution includes two stages: Depthwise and Pointwise Convolution, and this setup decreases the number of parameters 8 to 9 times compared to a standard convolution.

MobileNet-v2 innovated the Depthwise Separable Convolution structure into three sequential layers and integrated two new features: inverted residual connections, and linear bottlenecks. Experimental results in [4] show that to achieve the same classification or detection accuracy, MobileNet-v2 requires fewer parameters compared to the original MobileNet. Thus, this makes MobileNet-v2 the best option to integrate a detection model on a limited hardware configuration.

#### 2) Single Shot Detector (SSD)

This method was introduced with the expectation to be a Deep Learning architecture for vision applications embedded on mobile devices. The SSD [5] model predicts the probabilities of every class along with their corresponding bounding boxes' coordinates. This approach is very similar to YOLO [10] and they are both belong to a detection method called one-stage detector. Contrast to two-stage detectors, such as Faster R-CNN [11], one-stage detectors achieve much faster computational speed but still have competitively accuracy. One-stage detectors don't require an external Region Proposal stage after the feature extracting process, like two-stage detector, in order to correctly function, therefore make it very fast.

### B. Visual Human Tracking

Since Deep-learning based detection methods usually consume a lot of computing power, we used a light-weight

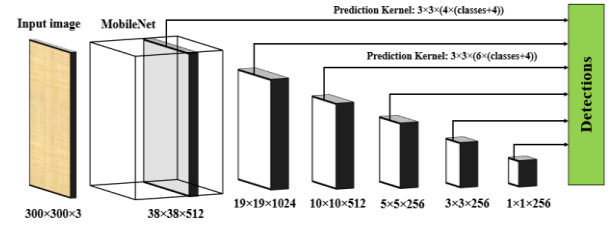


Fig. 3. MobileNetv2-SSD architecture

visual tracking algorithm to work along with it. This setup is great for object tracking applications in general because of the two methods' advantages: the detecting accuracy of the Deep-learning model and the fast computing speed of the visual tracking's correlating operation.

To increase the robustness of the detecting task, a visual tracking technique based on correlation operation was integrated. Moreover, using visual tracking can preserve a lot of computational power for the main computer, since Deep Learning models usually consumes a lot of computing memory especially on Embedded Hardware like a Raspberry Pi. This tracking method was built using [12], and it calculates the next position of the object by correlating the filter with the next frame. Figure 3 illustrates the working sequence of the technique. The tracking filter is initially created using the detected area of the human object in the first frame. The person is tracked by correlating the tracking filter on the next frame; the maximum value in the correlation response map indicates the new position of the person. The online update for the tracking filter is then performed based on the new location.

Compared to other tracking methods, tracking filters created by MOSSE [12] are better at differentiating between tracking targets and background, and more robust to changes in targets' appearance. To create a tracking filter, MOSSE first needs a set of training images and desired output responses. The desired output  $y_{desired}$  is created to have a 2D Gaussian distribution peak centered on the targets' position in training samples (Figure 3). The calculated output  $Y$  from input tracking region  $X$  is called the response map, which represented how similar is the tracking regions between two consecutive frames. If this response signal is weak (not a strong 2D Gaussian peak), the tracking process will fail. Note that, correlating calculation is carried out in the Fourier Domain, due to the element-wise multiplication of 2 matrices.

$$Y = X \odot F^* \quad (1)$$

The MOSSE creates and updates a filter  $F$  that minimizes the Sum Square Error between the actual response map and the desired one. This updating task is conducted on the whole training dataset and also while we are tracking people.

$$\min \sum_i |Y_i - Y_{desired}|^2 = \min \sum_i |X_i \odot F^* - Y_{desired}|^2 \quad (2)$$

To minimize the Equation 2, the derivative of Equation 2 with respect to  $F^*$  must equal to zero:

$$\frac{\partial}{\partial F^*} \sum_i |X_i \odot F^* - Y_{desired}|^2 = 0 \quad (3)$$

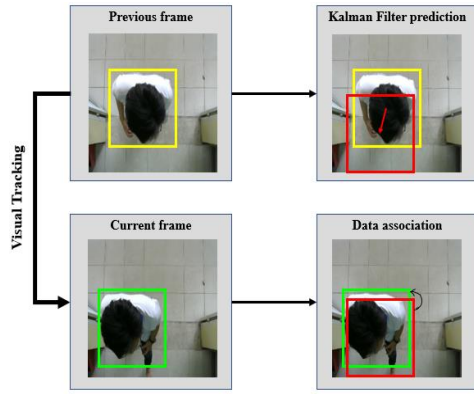


Fig. 4. ID tracking process

$$IoU = \frac{\text{Overlap Area}}{\text{Union Area}}$$

Fig. 5. Intersection over Union

Solve for  $F^*$  using Equation 3, the expression of MOSSE is found and represented in Equation 4:

$$F^* = \frac{\sum_i Y_{desired} \odot X_i^*}{\sum_i X_i \odot X_i^*} \quad (4)$$

### C. Data association method

In order to count the number of people, we need to know whether they have passed the entrance or not. This work cannot be merely done by detection, but require a method to associate the current detected data, and the next one. In this project, we used the SORT algorithm, introduced by [6] and it has proved to be very efficient when tracking multiple objects with small occlusion.

SORT used a Kalman filter for object tracking with the correction information from associating objects in two adjacent frames. The method used for the association between the same object was Intersection over Union (IoU). Figure 4 demonstrates the working principle of SORT, and Figure 5 shows the Intersection over Union calculation. To correctly assign the identity of two consecutive detections, the algorithm generates a prediction of the next position on the first detection, then uses the IoU criteria (IoU threshold of 0.3) to compare the prediction with the second detection.

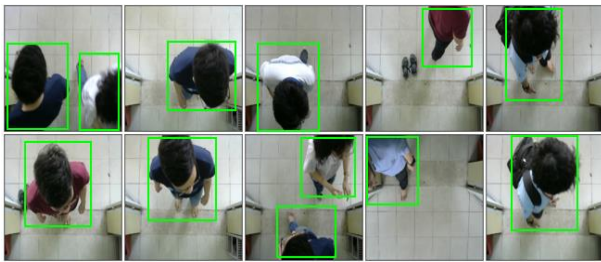


Fig. 6. Samples of training dataset

## III. EXPERIMENTS

### A. Training and Testing of MobileNet2-SDD

The dataset was divided into two subsets: a set of 1500 training images, a set of 500 validating images. Some examples of training data are shown in Figure 6. The program was written using Tensorflow Object Detection API and the training process was implemented using Google Colaboratory's GPU. Originally, the MobileNet2-SSD model was pre-trained on the COCO dataset, which has human as one of its detection classes. Fine-tuning a pre-trained model, especially the one that was trained on similar dataset, helps the training task to become simpler and faster.

The model was fine-tuned with small batches of 24 samples for each step, with the learning rate of 0.004. We used a much lower learning rate compared to the original training configuration of MobileNet2-SSD because lower learning rate can help the model to better generalize our custom dataset. The loss function for classification and localization is Weighted\_sigmoid and Weighted\_smooth\_L1. The model was fine-tuned around 10 hours and the total loss was approximately 1.60 as illustrated in Figure 7. The total loss diagram has been smoothed in order to show the decreasing tendency over training steps.

The model's accuracy was measured using the COCO evaluation metric. Figure 8 demonstrates the model's accuracy, which was calculated every 1000 training steps. As we could see from Figure 8, the pre-trained model performed well on our custom dataset (mAP was approximately 70%, and after 10 hours of fine-tuning, a mAP score of 83% was achieved. We stopped the training process when there was a sign of saturation in mAP [13] score, which indicated that the model's accuracy could not go any higher.

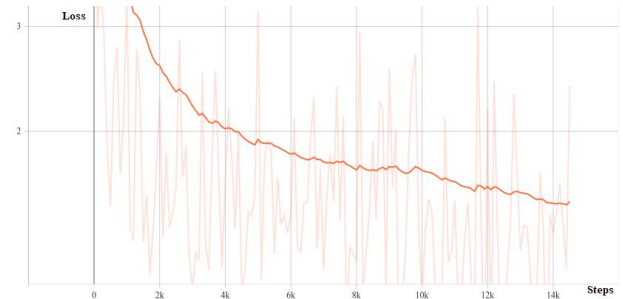


Fig. 7. Total loss diagram over the training process

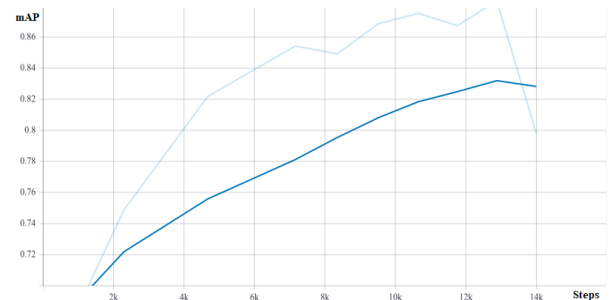


Fig. 8. Accuracy diagram over the training process

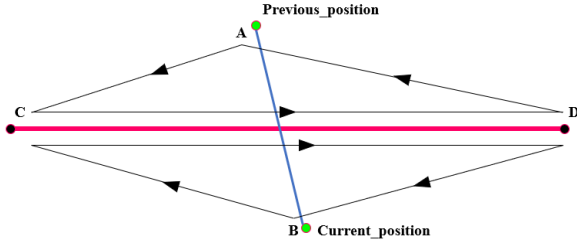
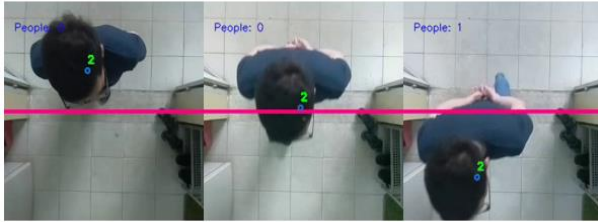


Fig. 10. Line intersection checking method

### B. Testing of the people-counting system

To count the people going in and out, we used a line intersection checking method. Figure 9 demonstrates the principle of the method. The line CD is our reference line, to determine whether a person has passed or not. A and B are the previous and current detected position of two consecutive frames. Whenever, the ACD and BCD form an opposite circular direction, AB and CD will intersect.

Some people counting results are provided in Figure 10. In the first scene (Figure 10a), there is one person coming into the room. The second scene (Figure 10b) is two people walking in at the same time. The third one (Figure 10c) is in different background setup. The system operated real-time on a Raspberry Pi 3B+ and achieved 7.5 frames per second (calculated for every 100 frames). The only limitation was the expensive people detection phase, which could cause misdetection for the system due to the long processing time of the Raspberry Pi.



(a) The result when one person coming in



(b) The result when two people coming in



(c) The result in different setup

Fig. 9. Sample results of people counting system

## IV. CONCLUSION

In this research, a people counting system running real-time on a Raspberry Pi was proposed. This work consists of three parts: people detection, tracking, and data association. To detect people, we used the Convolution Neural Networks on input images. The architecture of the detection model is MobileNetv2-SDD, which was fine-tuned on our custom dataset. From the detected areas of every person within the captured images, a filter was initialized and then correlated on the next image sequence to track detected people's position. Data association was used to create a tracking path of every person, from that the counting algorithm was implemented. The proposed system allow us to count people moving in and out of the entrance door. Experimental results demonstrates that our proposed system has the ability to accurately count people in different illuminating conditions and scenes, but still achieves a fast overall processing speed on a computing-constrained hardware. In the future, it would be worthwhile to decrease the processing speed of the detection phase to further improve our system.

## REFERENCES

- [1] D. Lefloch, F. A. Cheikh, J. Y. Hardeberg, P. Gouton, and R. Picot-Clemente, "Real-time people counting system using a single video camera," in *Real-Time Image Processing 2008*, International Society for Optics and Photonics. SPIE, 2008, pp. 71–82.
- [2] B. Li, J. Zhang, Z. Zhang, and Y. Xu, "A people counting method based on head detection and tracking," in *2014 International Conference on Smart Computing*, 2014, pp. 136–141.
- [3] J. García, A. Gardel, I. Bravo, J. L. Lázaro, M. Martínez, and D. Rodríguez, "Directional people counter based on head tracking," in *IEEE Transactions on Industrial Electronics*, 2013, pp. 3991–4000.
- [4] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018.
- [5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015.
- [6] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *CoRR*, vol. abs/1602.00763, 2016.
- [7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.
- [10] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015.
- [11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.
- [12] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2544–2550.
- [13] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: a retrospective," *Int. J. Comput. Vision*, vol. 111, 2015, pp. 98–136.



# Electricity Demand Forecasting for Smart Grid Based on Deep Learning Approach

Van-Binh Nguyen

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
nguyenvanbinh120989@gmail.com

Minh-Thien Duong

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
duongthien2206@gmail.com

My-Ha Le<sup>✉</sup>

Faculty of Electrical and Electronics  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
halm@hcmute.edu.vn

**Abstract**—Electric power load demand forecasting plays a very indispensable role in energy management, such as suitable planning and investing in developing more infrastructure, and moderate operating the electricity system. Moreover, the accuracy of the electricity demand forecasting must be paid special attention. Although research on this problem has been conducted in recent years, the authors noticed that elevated accuracy outcomes have not been gained yet and have provoked controversy about forecasting results. In this paper, an electricity demand forecasting method based on a Deep Learning model, namely Long-Short-Term-Memory (LSTM), which is an improvement of Recurrent Neural Networks (RNNs) is proposed. The proposed network architecture includes four layers: sequence input, LSTM, fully connected, and regression output. The proposed method was implemented on the power consumption data in six years, from 2012 to 2017, of Tien Giang province, Vietnam. The forecasting result, evaluated using Root Mean Square Error (RMSE), was 9.63, suggesting that this would be a great contribution to studies in the energy sector.

**Keywords**—Electrical Load Forecasting, Smart Grid, Deep Learning, Long-Short-Term-Memory (LSTM), Neural Network.

## I. INTRODUCTION

In the energy management industry, forecasting the electric load demand is extremely important because it is closely linked and directly affects the daily lives of people and economic sectors. In addition, load demand forecasting is crucial in ensuring the working safe mode and energy economical for the electricity system. At the same time, it plays a paramount importance role in planning the system development strategy. In this context, electric load forecasting is not only for the detection of system instabilities and protection [1], [2] but also for effective energy management [3]. Because the electrical load is primarily a univariate time series [4], many time series forecasting methods can be put into practice for electrical power load forecasting. Conventional load forecasting methods can often not fully and accurately describe the actual process that occurs, because the number of databases is incomplete, there are many errors or it takes a long time for calculations. In fact, there are no equations with available parameters but only approximate values or mathematical expectations. Therefore, an existing equation must be given with unknown parameters, then an approximate method to find these parameters is used, which will reduce the accuracy remarkably. The traditional methods are used effectively only in cases where the data are linearly related to each other. It is not possible to clearly show the complex, nonlinear relationships between the load and the relevant parameters. To overcome the disadvantages of

traditional load forecasting methods, scientists have applied modern forecasting techniques such as neural networks, fuzzy logic, regression...[5], [6], [7], [8], [9], [10]. The forecasting methods aforementioned are increasingly concerned because the forecast results are quite accurate. If the forecast is too much compared to the demand, the mobilization will be too large. As a result, there will be an increase in investment capital, which can possibly cause energy losses. On the contrary, if the load forecast is too low compared to the demand, there will not be enough supply of energy. This will lead to the removal of some loads without prior planning, which may damage the economy. Moreover, many techniques for forecasting the power load have been carried out before such as mathematical methods [11], [12] statistical algorithms [13], and especially, methods related to artificial intelligence. Recently, the artificial neural network has achieved many outstanding results because it is effective, and easy-to-implement. Therefore, electric load demand forecasting using artificial neural network technology is studied in this paper.

The rest of the paper is organized as follows. In Section 2, we present the proposed method of electric load demand forecasting. Experimental results in Section 3 demonstrate the performance of the proposed approach. Section 4 is the conclusion of the paper.

## II. PROPOSED METHOD

### A. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a type of deep learning of artificial neural network, developed in the 1980s. The main idea of RNNs is to use sequential information independently of each other. RNNs are called “recurrent” because they enforce the same assignment for all elements of a sequence with the output contingent on previous computations. In other words, RNNs have the ability to remember previously computed information. In theory, RNNs can use a piece of very long information, but in practice, it can only remember a few steps before [14].

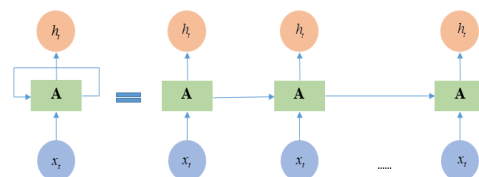


Fig. 1. The description of the RNNs model

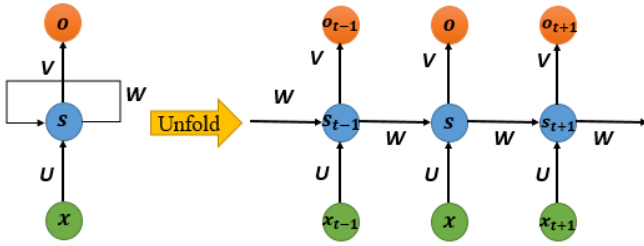


Fig. 2. The math description of the RNNs model

RNNs is a network that includes many identical neural nodes, each is directly connected to every other one. Unlike a traditional deep neural network, which uses different parameters at each layer, an RNNs ability to share the same parameters through whole steps. This significantly decreases the total number of parameters that need to learn. Each connection has an adjustable real weight. Input nodes receive data from outside the network, output nodes are the final result, hidden nodes adjust data on the way from input to output (Figure 1). The details of an RNNs are depicted in Figure 2. In that,  $x_t$ ,  $s_t$ ,  $o_t$  are input, the hidden state, and output at step  $t$ , respectively.

#### B. Long-Short-Term-Memory (LSTM Networks)

Long-Short-Term-Memory (LSTM Networks) are popular models that have shown the potential in many forecasting tasks [15]. LSTM is designed to avoid long-term dependencies issues [16]. Memorizing information over the long term is an advantage of this network. All RNNs have the outline of a chain of iterative modules of neural networks. In standard RNNs, this iterative module will have a super simple architecture, such as a single hyperbolic tangent layer (Figure 3). LSTM also has this chain-like structure, but the iterative module has various structures. Instead of having a single neural network layer that has a lot of interactions in a very particular modality (Figure 4).

The core of LSTM is the cell state rephrased by the horizontal line running across the top of the schema. The cell state is like a conveyor belt. It runs across the nodes with only some minor linear interactions. So that the information can easily be transmitted smoothly without fear of being changed (Figure 5). The LSTM does have the capability to remove or add information to the cell state, carefully regulated by structures denoted gates (Figure 6). Gates are where the information screening passes through. They are constructed of a sigmoid neural network layer and a point-wise multiplication operation. The output of the sigmoid layer are numbers between zero and one. A value of zero means that no information is transmitted, contrary if it is one, that means all information passes through it.

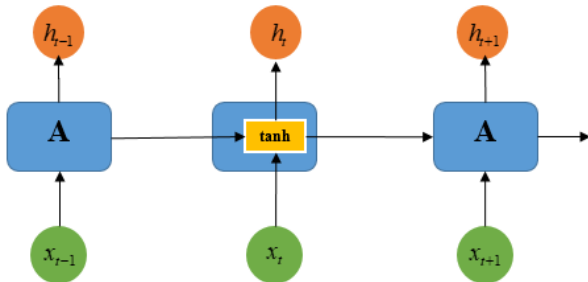


Fig. 3. The iterative module in a standard RNNs includes a single layer.

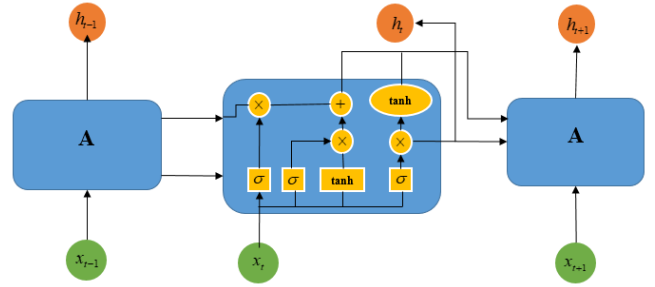


Fig. 4. The iterative module in an LSTM includes four interacting layers.

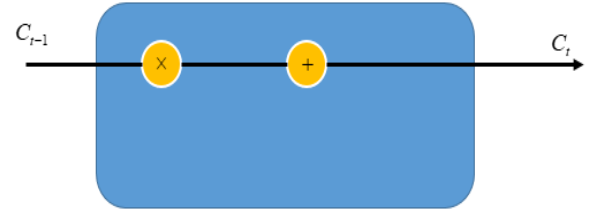


Fig. 5. The cell state of LSTM model.

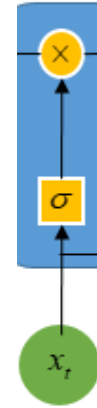


Fig. 6. The cell gate of LSTM model.

#### C. Math fundamentals

To begin with, LSTM is to decide which information to remove from the cell state. This decision is made by a sigmoid layer also known as the *forget gate cell*:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (1)$$

The next step is to decide which new information to save to the cell state by input gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (4)$$

Then, the old cell state ( $C_{t-1}$ ) is updated to the new state  $C_t$ :

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

In the final step, the output gate decides what the output is by associating the input and memory:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$



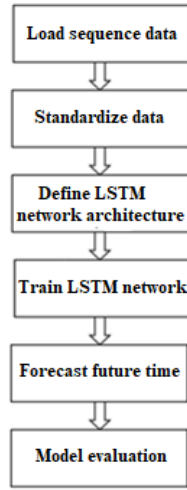


Fig. 7. The proposed flowchart for electrical load forecasting process.

### III. EXPERIMENT

#### A. Dataset

The data set is from the power load consumption in six years of Tien Giang province, Vietnam from 2012 to 2017, from this dataset, it is possible to forecast the electricity demand in the future. The amount of data is about 2200 values. The measurement unit is kWh.

#### B. Forecasting result

Our proposal model includes three main stages: The first one is loading sequence data. The second stage is to standardize the data. Subsequently, the authors define LSTM network architecture and train the LSTM network. After this phase completes, electricity demand in the future time will be forecasted. In the last stage of this process, we compare the predicted value with the test value to evaluate the model using RMSE. The flowchart of the implementation process of our model is illustrated in Figure 7.

a) Load sequence data: Load sequence data as the load value in a single time series with time steps corresponding to days and values corresponding to the power consumption. The output is a cell array, where each constituent is a single time step (Figure 8). From the database, the authors divide them into two parts separately. One is the training data and another is test data with an 80:20 proportion.

b) Standardize data: To reduce input data's noise as well as to prevent divergence during training, standardization is extremely important. Specifically, we standardized the training data and test data to have zero mean and unit variance (Figure 9). For a random variable vector  $x$  made up of  $N$  scalar observations, the standard deviation is calculated to follow formulation (8), where  $\mu$  is the mean of  $x$  as (9), a value is standardized as (10).

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^N |x_i - \mu|^2} \quad (8)$$

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (9)$$

$$y = \frac{x_i - \mu}{S} \quad (10)$$

c) Define LSTM network architecture: The four layers of the proposed LSTM regression network structure are

Sequence Input, LSTM, Fully Connected, and Regression Output. Specify the LSTM layer have 250 hidden units (Figure 10).

d) Train LSTM network: The training process is set up with 200 epochs and the gradient threshold to 1. The initial learn rate 0.005 and drop the learning rate after 125 epochs by multiplying by a factor of 0.2. The authors took advantage of the NVIDIA GeForce GTX 1080 Ti GPU with 12GB of memory, CUDA Cores 3584 for network training. The total time that the authors conducted the training process in conjunction with 200 epochs for 1760 samples was approximately 100 minutes (Figure 11).

e) Forecast future time: The result of electricity demand forecasting of Tien Giang province, Vietnam is depicted in Figure 12.

f) Model evaluation: The result after forecasting will be compared to the test value in the original dataset. Looking at the Figure 13 more closely, one can see that these two lines fit together, this demonstrates that the forecast results of the electricity demand are relatively accurate. Root Mean Square Error is selected as the error metric. The use of RMSE is a very common error metric for numerical predictions. RMSE is computed as the square of the correlation between the observed  $y$  values and the predicted  $\hat{y}$  values as (11), namely  $RMSE=9.6333$ .

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (11)$$

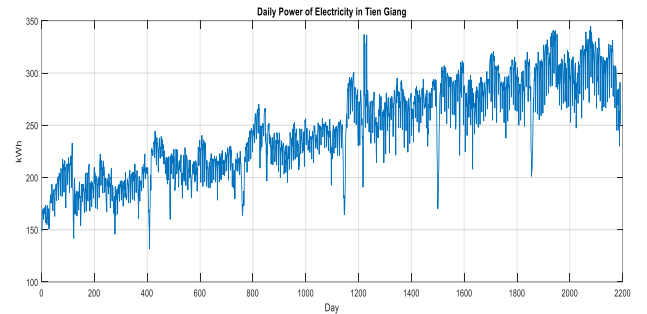


Fig. 8. The daily power data of electricity in Tien Giang province, Vietnam from 2012 to 2017.

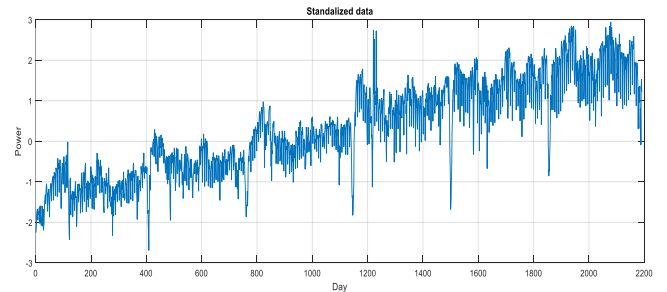


Fig. 9. The daily power data of electricity in Tien Giang province, Vietnam from 2012 to 2017 after standardize process.

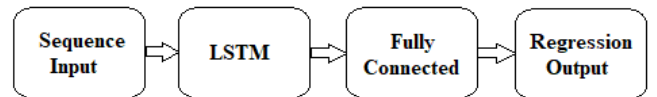


Fig. 10. The block diagram illustrates the architecture of the proposed LSTM network.

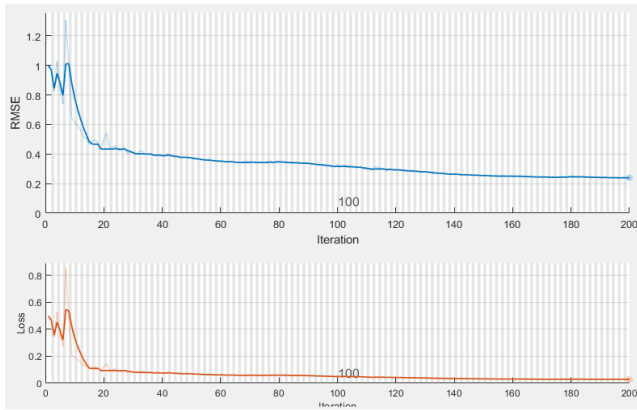


Fig. 11. The training process of the proposed forecasting model.

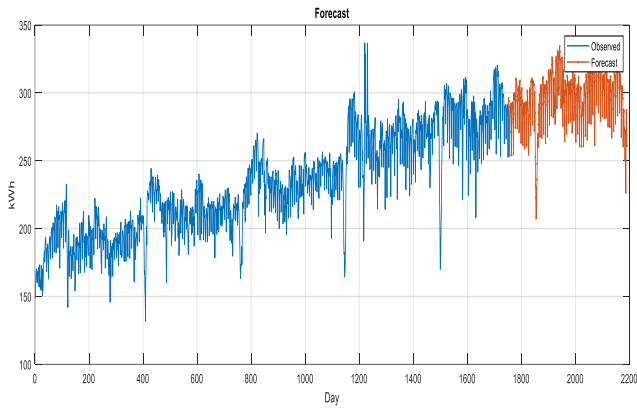


Fig. 12. The electricity demand forecasting results in the future time of Tien Giang province, Vietnam.

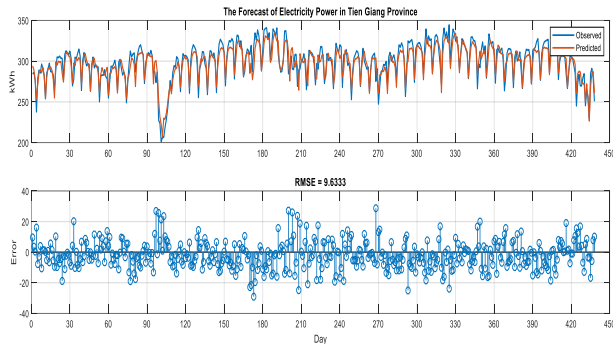


Fig. 13. Model evaluation by comparing the predicted value with the test value.

### C. Comparison with other solution

To emphasize the outstanding results of the proposed forecasting method, the authors have experimented with a traditional neural network to forecast load demand based on the same data set. It has been proven that the results of our proposed method are more surpassing than the results of the method using the traditional neural network. Training process diagrams have shown that test results are diverging (Figure 14). This demonstrates that the forecast results are only true at the beginning values, the error of the forecasting results increases gradually from the end values. That is why forecasting results are instability across time. According to figure 15, we can see that when using the traditional neural network, the forecasting results do not fit the test values from the original dataset. This gives rise to the forecast of the load demand which is not highly accurate.

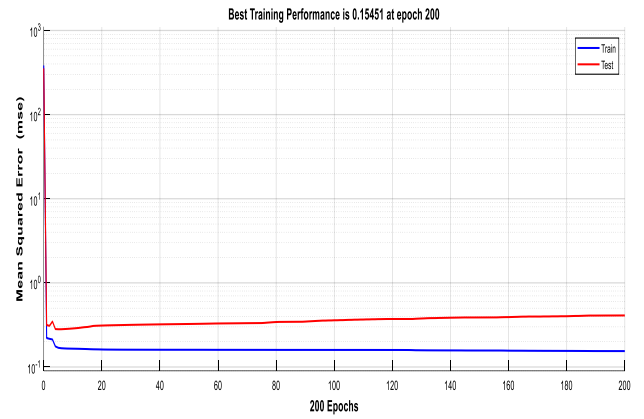


Fig. 14. Training process of traditional neural network.

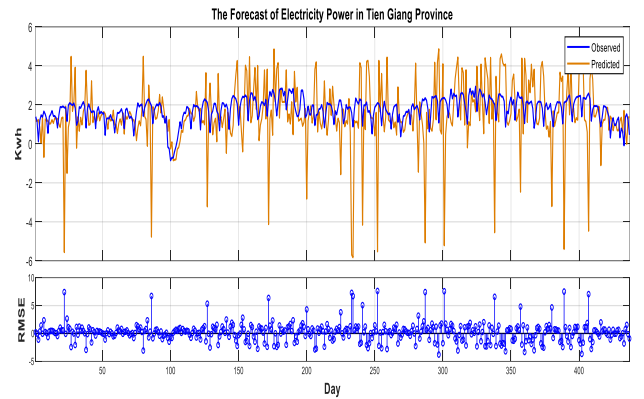


Fig. 15. Model evaluation by comparing the predicted value with the test value when using traditional neural network.

## IV. CONCLUSION

Experimental results have shown that Long-Short-Term-Memory networks can solve the disadvantages that traditional networks have not achieved yet. In this paper, the authors forecasted the electric load demand of the real dataset. The authors had analyzed the theory and implemented a forecast on simulation software. The proposed method using the LSTM network to forecast has higher accuracy than the previous traditional methods. To sum up, this research has obtained accurate forecasting results of electricity demand. Therefore, this research result was applicable to reality in electricity agencies nationwide.

Although the results obtained from the study are satisfied, there are still several issues that the authors need to improve. For deep learning applications, data need to be collected on a larger scale to acquire more accurate results. In addition, adjusting the network architecture for more optimal results is also a problem that the authors are interested in.

### ACKNOWLEDGMENT

This research was implemented at Intelligence Systems Laboratory (ISLab), Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology and Education, Vietnam. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which significantly improved the quality of this paper.

### AUTHOR CONTRIBUTIONS

These authors contributed equally to this work.

## REFERENCES

- [1] K. Dehghanpour, Z. Wang, J. Wang, Y. Yuan and F. Bu, "A Survey on State Estimation Techniques and Challenges in Smart Distribution Systems," in *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2312-2322, March 2019, doi: 10.1109/TSG.2018.2870600.
- [2] G. B. Giannakis, V. Kekatos, N. Gatsis, S. Kim, H. Zhu and B. F. Wollenberg, "Monitoring and Optimization for Power Grids: A Signal Processing Perspective," in *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 107-128, Sept. 2013, doi: 10.1109/MSP.2013.2245726.
- [3] L. Zhang, V. Kekatos and G. B. Giannakis, "Scalable Electric Vehicle Charging Protocols," in *IEEE Transactions on Power Systems*, vol. 32, no.2, pp. 1451-1462, March 2017, doi:10.1109/TPWRS.2016.2582903.
- [4] E. Almeshai and H. Soltan, "A methodology for electric power load forecasting," *Alexandria Engineering Journal*, vol. 50, no. 2, pp. 137-144, 2011, doi: 10.1016/j.aej.2011.01.015.
- [5] W. Charytoniuk and M.-S. Chen, "Very short-term load forecasting using artificial neural networks," *IEEE Trans. Power Systems*, vol. 15, no. 1, pp. 263-268, Feb. 2000.
- [6] Dong-Xiao Niu, Qiang Wanq and Jin-Chao Li, "Short term load forecasting model using support vector machine based on artificial neural network," *2005 International Conference on Machine Learning and Cybernetics*, Guangzhou, China, 2005, pp. 4260-4265 Vol. 7, doi: 10.1109/ICMLC.2005.1527685.
- [7] Z. Yun, Z. Quan, S. Caixin, L. Shaolan, L. Yuming, and S. Yang, "RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 853-858, Aug. 2008, doi: 10.1109/TPWRS.2008.922249.
- [8] LOU, Chin Wang; DONG, Ming Chui. Modeling data uncertainty on electric load forecasting based on Type-2 fuzzy logic set theory. *Engineering Applications of Artificial Intelligence*, 2012, 25.8: 1567-1576.
- [9] A. Khotanzad, Enwang Zhou and H. Elragal, "A neuro-fuzzy approach to short-term load forecasting in a price-sensitive environment," in *IEEE Transactions on Power Systems*, vol. 17, no. 4, pp. 1273-1282, Nov. 2002, doi: 10.1109/TPWRS.2002.804999.
- [10] M. Hassanzadeh and C. Y. Evrenosoğlu, "Power system state forecasting using regression analysis," *2012 IEEE Power and Energy Society General Meeting*, San Diego, CA, 2012, pp. 1-6, doi: 10.1109/PESGM.2012.6345595.
- [11] M. Hassanzadeh, C. Y. Evrenosoğlu and L. Mili, "A Short-Term Nodal Voltage Phasor Forecasting Method Using Temporal and Spatial Correlation," in *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3881-3890, Sept. 2016, doi: 10.1109/TPWRS.2015.2487419.
- [12] Wang Y., Niu D., Ji L. (2011) Optimization of Short-Term Load Forecasting Based on Fractal Theory. In: Nguyen N.T., Trawiński B., Jung J.J. (eds) *New Challenges for Intelligent Information and Database Systems. Studies in Computational Intelligence*, vol 351. Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-19953-0\_18.
- [13] Y. Chakhchoukh, P. Panciatici and L. Mili, "Electric Load Forecasting Based on Statistical Robust Methods," in *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 982-991, Aug. 2011, doi: 10.1109/TPWRS.2010.2080325.
- [14] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.
- [15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104-3112.
- [16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

# A Novel Technique for Increasing Concentration Ratio and Uniformity of Fresnel Lens

Thanh-Tuan Pham\*

Renewable Energy Department  
Faculty of Vehicle and Energy  
University of Technology and Education  
Ho Chi Minh City, 71307, Viet Nam  
tuanpt@hcmute.edu.vn

Thanh Phuong Nguyen

Faculty of Graphic Arts and Media  
University of Technology and Education  
Ho Chi Minh City, 71307, Viet Nam  
phuongnt@hcmute.edu.vn

**Abstract**—In this paper, we propose a technique using the conservation of optical path length and edge ray theorem to design a Fresnel lens, which can achieve high concentration ratio, high uniform irradiance, and small f-number. The structure of the Fresnel lens consists of three parts: Inner part, middle part, and outer part. The design process is carried out by solving the equations of the conservation of optical path length in Matlab<sup>TM</sup> program. In addition, a wide range of wavelengths is applied to do rays tracing so that the system becomes more suitable in real conditions. In this technique, the Fresnel lens is constructed by many grooves that are built by the ideal Cartesian oval surface. Thus, the optical efficiency of the designed lens is improved. The simulation results by LightTools<sup>TM</sup> software show that the Fresnel lens has good optical properties such as a high concentration ratio of 900x, f-number = 0.46, high uniform irradiance distribution, and optical efficiency larger than 85%.

**Index Terms**—Fresnel lens, Concentrator photovoltaic (CPV), high concentration ratio, uniform irradiance distribution.

## I. INTRODUCTION

Recently, the development of photovoltaic technology improves efficiency of solar cell significantly. The latest record of direct conversion of sunlight into electricity of multi-junction solar cell reaches 46% published by Fraunhofer, Germany [1]. Therefore, solar energy becomes a promising resource to replace completely fossil energy in the future [2]–[4]. However, the cost of multi-junction cells is high which leads to high price of a photovoltaic system. An effective way to reduce the cost is cutting down the amount of the required cell area while increasing the cell efficiency by using concentrator photovoltaic system (CPV) [2], [5]. CPV is a photovoltaic technology that generates electricity from sunlight by using lenses or mirrors to focus sunlight to the multi-junction (MJ) solar cells.

In a CPV, design of the concentrator is an important task to achieve high performance. Fresnel lens [6]–[8] has been used widely as a concentrator in CPV because it has light weight, mass production, and uses cheap materials [9]. Nevertheless, some disadvantages of these concentrators are long focal distance and non-uniform irradiance. The long focal distance increases the volume and weight of CPV systems. Non-uniform irradiance that makes a hot spot points degrades the reliability, the conversion efficiency, and the lifetime of

solar cells [9], [10]. Therefore, concentrators of CPV have been usually designed by using the primary optical element (POE) and the second optical element (SOE) [11]–[13]. The primary element, Fresnel lens, is used to collect and guide the sunlight to an area of SOE. The SOE, which is placed closely to the cell surface, is used to redistribute the sunlight uniformly over the solar cell. Hence, the design process becomes more complicated and increases the cost of CPV system in these structures [7]. To overcome this problem, there are some efforts to design CPV system without using SOE such as Juan. C. Gonzalex [7], Kwangsu Ryu [14], Jui-Wen Pan [9], etc. In these design methods, Fresnel lens was modified to achieve uniform irradiance over the solar cell surface. However, they still have some limitations in terms of increasing concentration ratio and uniformity of irradiance distribution.

In this paper, we introduce an optical design technique to achieve Fresnel lens with high concentration ratio, uniform irradiance, optical efficiency, and small f-number. The designed Fresnel lens consists of three parts: inner, middle, and outer areas. The inner part is built by using refraction phenomena, whereas the middle and outer area are constructed by using total internal reflection (TIR). The designed Fresnel lens can be used to manufacture CPV without using secondary optical element. In this technique, the shape of Fresnel lens is constructed by many grooves, which are built by Ideal Cartesian Oval surfaces [16]. The design process is conducted by using Matlab. Ray tracing technique in LightTools<sup>TM</sup> software is used to optimize the structure of lens. The shape of lens is drawn in 2-D and 3-D in LightTools<sup>TM</sup> as well.

## II. DESIGN METHOD

The design process is carried out with three parts named as inner area, middle area, and outer area in Matlab program. The design flow chart of optimum lens shape is illustrated in Fig. 1. The process is performed gradually following diameter of lens from center to the edge of the lens. Firstly, the program constructs the inner area limited by the position where the total internal reflection (TIR) appears. Because of the total internal reflection (TIR), the ray coming to the lens cannot guide to the receiver so that the design method has to be modified for the middle area. Finally, the design method is changed one more

time for the outer area to help increase the size of the Fresnel lens while the F-number keep no change. The estimation of limitation among areas and the way to design for three areas are described in detail as following parts.

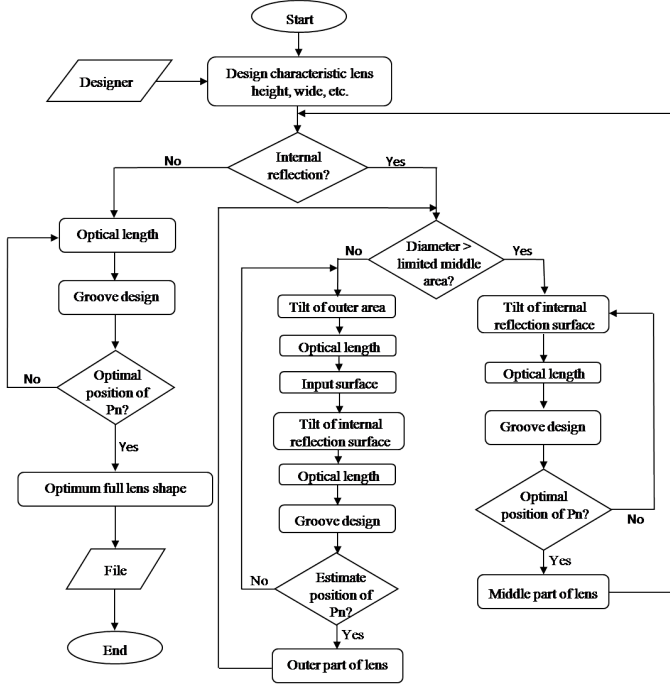


Fig. 1. The flow chart of optimum calculation in Matlab program.

### A. Design of Inner Area

The inner area of designed Fresnel lens is designed by using refraction phenomena. The bundle of direct sunlight coming to each groove, refracts at exit surface. Then, it focuses at the  $P_{ng}$  focal point of each groove. After that it is distributed uniformly over receiver that is demonstrated in Fig. 2 (the picture is not scale). The edge ray, which comes to the left side of groove, refracts at the exit surface (position  $P_s$ ) then it goes forward right side of receiver (position  $x_2$ ). In contrast, the edge ray, which comes to the right side of groove, refracts at the exit surface (position  $P_n$ ) to go to the left side of receiver (position  $x_1$ ). The left and right edge rays intersect each other at focal point  $P_{ng}$ . Every ray between the left and right edge ray will be focused at  $P_{ng}$  and then reach somewhere at the receiver. Therefore, the bundle of direct sunlight coming to one groove is distributed uniformly over receiver.

In this design, the irradiance distribution, which can be non-uniform or uniform, depends on position of focal point  $P_{ng}$ . If the position of focal point is estimated wrong the irradiance distribution can be shifted to left, right, or both sides of the receiver. The focal point  $P_{ng}$  can be estimated by using the conservation of optical path length as follows.

$$a_1 = na_2 + a'_2 = na_m + a'_m = na_{end} + a'_{end} = OPL \quad (1)$$

Where  $OPL$  is the optical path length constant,  $n$  is the refractive index.  $a_1$ ,  $a_2$ ,  $a'_2$ ,  $a_m$ ,  $a'_m$ ,  $a_{end}$ , and  $a'_{end}$  are the

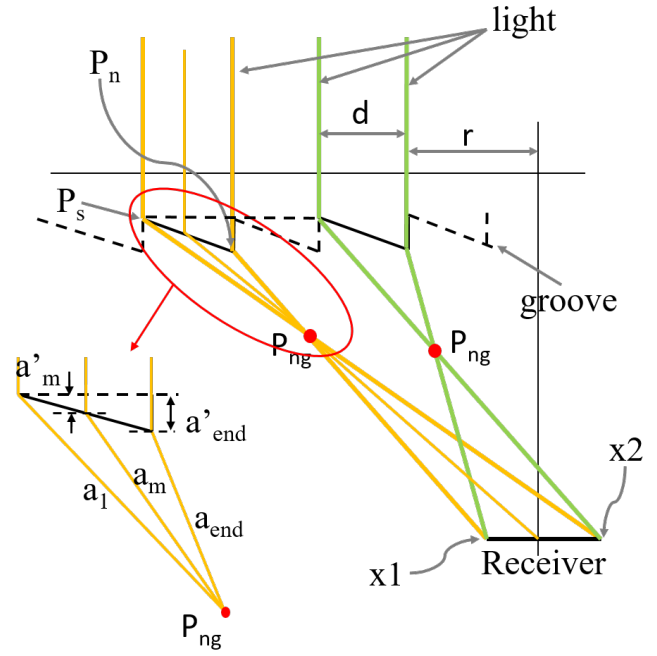


Fig. 2. The distribution of sunlight beam over receiver by TU-Fresnel lens.

path lengths of rays in corresponding medium. Fig. 2 shows all of those elements. Limitation of inner area is determined by appearance of total internal reflection (TIR) (see Fig. 3). The bigger of the diameter of the lens is, the wider of the refraction angle is until  $90^\circ$  at which the TIR appears. For this reason, the rays come to the groove of lens cannot reach to the receiver. Therefore, The boundary of the inner area can be calculated by using Snell's law and TIR phenomena. The limitation of inner area's diameter is estimated through  $\beta$  that is illustrated in Fig. 4.

$$\alpha + \beta = 90^\circ \quad (2)$$

$$n \sin \alpha = \sin \alpha + \beta \quad (3)$$

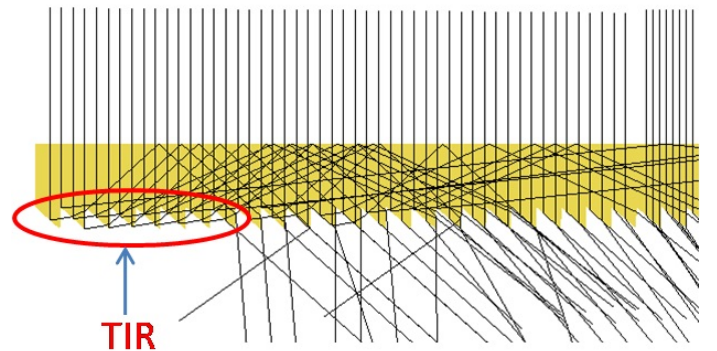


Fig. 3. The total internal reflection appears at grooves of lens by ray tracing.



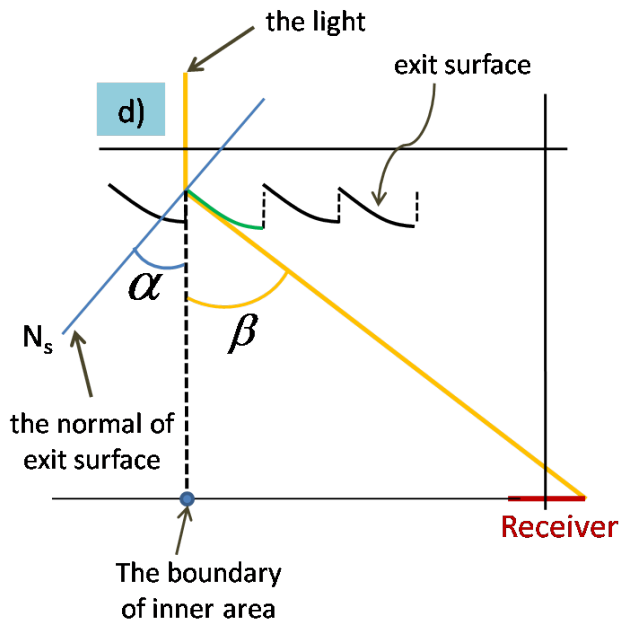


Fig. 4. The inner area boundary is estimated using TIR phenomena.

There is a limitation for increasing the diameter (concentration ratio) of Fresnel lens because of TIR phenomena. Therefore, the way of design process has to change from using refraction to using total internal reflection. The design technique using TIR used to construct the middle area of designed Fresnel lens is presented in detail in next section.

#### B. Design of Middle Area

The middle area of the Fresnel lens is designed by using total internal reflection phenomena. In this part, each grooves of lens consist of three surfaces: aperture surface ( $S_0$ ), TIR surface ( $S_1$ ), and exit surface ( $S_2$ ). Those are illustrated in Fig. 5. In each groove, the aperture surface (input surface) is a flat, which is vertical with the direct sunlight. Therefore, the rays pass through the aperture surface without any refraction. After that the rays come to The TIR surface, which is also flat. The TIR surface is used to reflect the light beam to the output surface, which is a Cartesian oval surface to guide the reflected rays to the focal point  $P_{ng}$  of the groove.

Similarly to the inner area, estimation of the position of the focal point  $P_{ng}$  is a key to distribute the light beam uniformly over receiver. The focal point  $P_{ng}$  is the intersection point of the left and the right edge rays of each groove. Some steps in the design procedure for one groove are described in detail as follows, with reference to Fig. 5.

- Step 1. The left edge ray reflects at  $P_0$  of  $S_1$  surface. Then it exits  $S_2$  surface at  $P_3$  point and reaches to the left side  $x1$  of receiver.
- Step 2. The point  $P_{ng}$  can be selected with the condition of being on the straight line joining points  $P_3$  and  $x1$ .
- Step 3. The right edge ray reflects at  $P_1$  of  $S_1$  surface. Then it exits the lens at  $P_2$  point of  $S_2$  surface and focuses

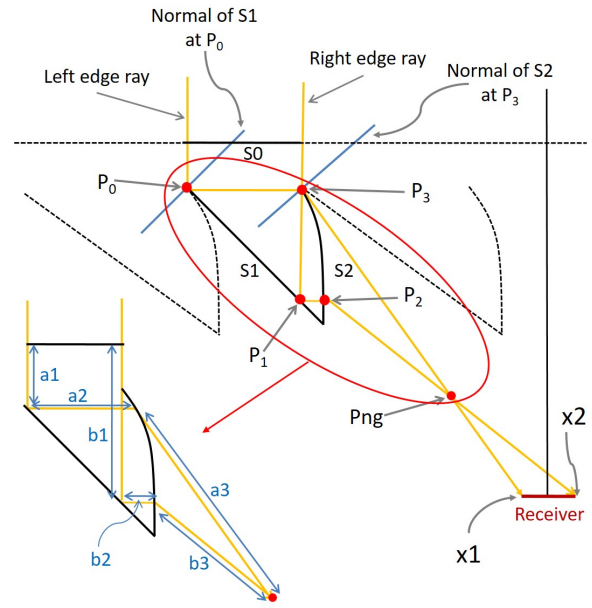


Fig. 5. The inner area boundary is estimated using TIR phenomena.

onto  $P_{ng}$  before coming to the receiver. If the right edge ray reaches to the receiver at a point, which is different from the  $x2$  position, the irradiance distribution does not fit on the receiver. Then the  $P_{ng}$  has to be chosen again (step 2).

- Step 4. When the focal point  $P_{ng}$  is estimated, the Cartesian oval surface  $S_2$  can be estimated by the requirement, in which in which all direct sunlight exiting the groove at  $S_2$  surface have to focus on the focal point  $P_{ng}$ .
- Step 5. The same procedure is repeated to design next groove.

As a result, the groove can be built and the light beam is distributed uniformly over receiver. In this procedure, the positions ( $P_0, P_1, P_2, P_3$ ) of one groove can be calculated by using the equation of optical path length following.

$$n(a_1 + a_2) + a_3 = n(b_1 + b_2) + b_3 = OPL \quad (4)$$

Where  $n$  is the refractive index of lens.  $a_1, a_2, a_3, b_1, b_2,$  and  $b_3$  are the optical path lengths in the medium. Fig. 5 indicates the optical path length in one groove of the middle area.

Lens is designed with optical loss as small as possible. In the middle area, the tilt of  $S_1$  surface can affect to optical loss of the grooves. Therefore, the tilt of  $S_1$  surface is investigated with some different angles. In Fig. 6, the total reflection angle, which is created by normal ( $N_s$ ) of  $S_1$  and reflected ray, is investigated with some values ( $\alpha = 45^\circ, \alpha = 50^\circ,$  and  $\alpha = 55^\circ$ ). The results in Fig. 6 show that the total reflection angle  $\alpha = 55^\circ$  is the best choice for the structure of groove in middle area.

In addition, the optical loss can be reduced more by changing the shape of groove. The rays, which come from the upper surface  $S_2$ , are interrupted by the surface  $S_1$  of next groove.

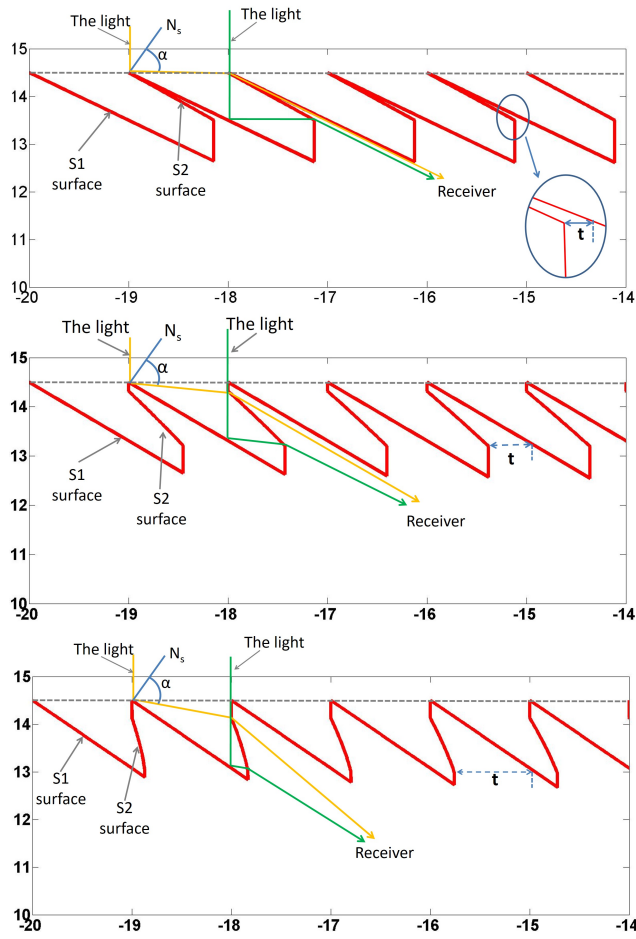


Fig. 6. The structured grooves in the middle area with different tilts of S1 surface a)  $\alpha = 45^\circ$ , b)  $\alpha = 50^\circ$ , c)  $\alpha = 55^\circ$ .

Hence, the tip of groove is cut to decrease this kind of optical loss as Fig. 7.

Dimension of the middle area is limited by groove's structure of the designed Fresnel lens. when diameter of the lens increases, the point  $P_2$  tends to move from right to left side of  $P_1$  point. Therefore, S2 surface tilts to the right side when  $P_2$  is left side of  $P_1$ . For this reason, the optical loss at S2 surface increases. Thus, limitation of middle area is estimated when  $P_2$  is asymptotic  $P_1$ . Fig. 8 demonstrates that  $P_0$ ,  $P_1$ ,  $P_2$ , and  $P_3$  are the critical points of one groove.

### C. Design of Outer Area

An effective way to increase concentration ratio (increase diameter of lens) while keeping small optical loss is addition of outer area (third part). Actually, the outer area is an extension of the middle area to overcome the limitation of the middle area. Thus, this part is designed in the same way as the middle area. The grooves of outer area consist of three surfaces: input surface ( $S_0$ ), TIR surface ( $S_1$ ), and exit surface ( $S_2$ ). In addition, the irradiance distribution process in one groove of this area is similar to the middle area. However, the height of the grooves decreases gradually in this part following a tilt

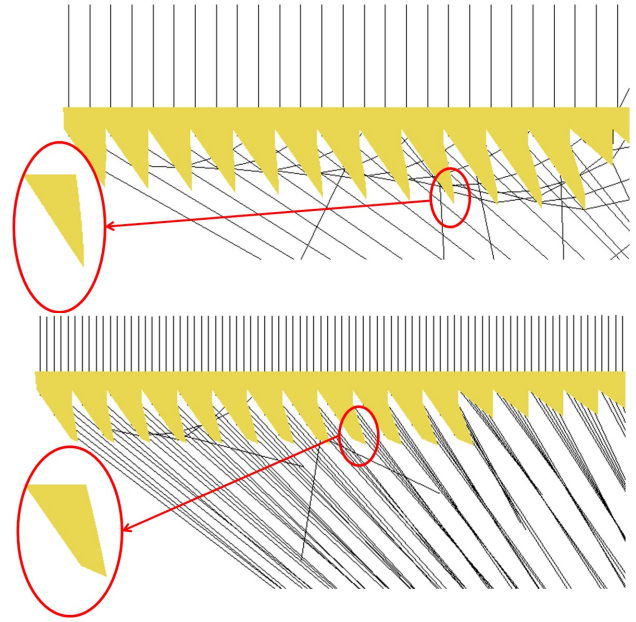


Fig. 7. The shape of a) the conventional groove and b) cutting groove to reduce optical loss.

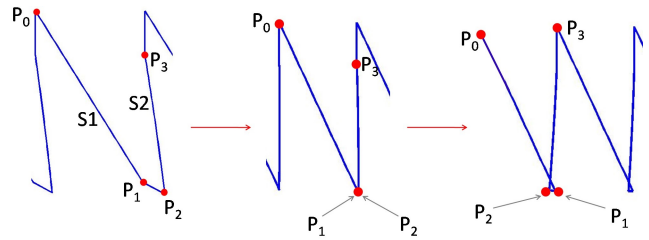


Fig. 8. The  $P_2$  position moves from right to left side of  $P_1$  when diameter of lens increases.

angle ( $\gamma$ ) while the height of the grooves in middle area is constant. Reducing the height of the grooves helps the rays to reach to the receiver. Hence, the optical loss is prevented in the lens when the diameter of the designed lens increase. The shape of grooves in the outer area is shown in Fig. 10.

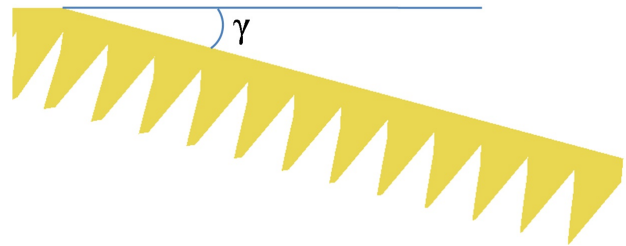


Fig. 9. The outer area is re-designed using the tilt angle  $\gamma$ .

In this design, the aperture surface ( $S_0$ ) is a tilt flat surface with angle  $\gamma$ . So, the light beam is refracted at  $S_0$  whereas there is no any refraction at  $S_0$  surface in the inner area and the middle area. Thus, the design process in outer area is little

more complicated than that of inner and middle. The light beam, which goes through one groove, can be described as in Fig. 10.

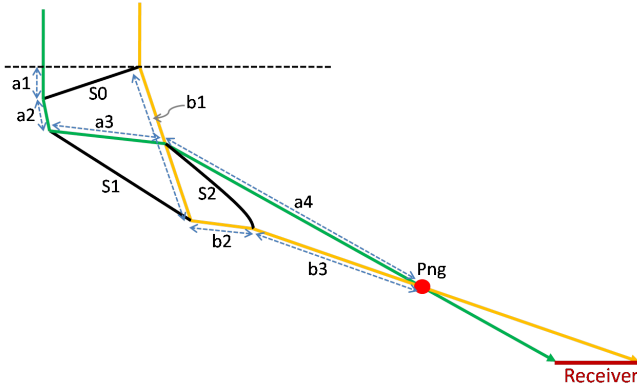


Fig. 10. The irradiance distribution of the groove, which has a tilt of aperture surface in outer area.

The irradiance distribution of the direct sunlight can be described gradually by using the conservation of optical path length (see Fig. 10). Firstly, the bundle of the direct sunlight, which comes to one groove in the outer area, refracts at aperture surface ( $S_0$ ). Secondly, it is total internal reflected at TIR surface ( $S_1$ ). Thirdly, it refracts again at exit surface. After that it focuses at  $P_{ng}$  position point. Finally, the bundle of rays distributes over the receiver. In this design process, the estimation of the position of focal point  $P_{ng}$ , which is a key to distribute irradiance uniformly over the receiver with any size, can be carried out by using (5) as follows.

$$n(a_1 + a_2 + a_3) + a_4 = n(b_1 + b_2) + b_3 = OPL \quad (5)$$

Where  $n$  is the refractive index of lens.  $a_1, a_2, a_3, a_4, b_1, b_2,$  and  $b_3$  are the optical path length of the left and right edge ray in the medium.

In this design, there are some advantages such as decreasing thickness of lens in outer area, reducing optical loss by material absorption, and cutting down weight of lens. However, the optical loss still appears in the outer area as shown in Fig. 11. The appearance of optical loss in this part can be explained by dimension of exit surface ( $S_2$ ). The  $S_2$  surface is so big that the rays at the top of exit surface are interrupted by the next groove.

This kind of optical loss can be reduced more by decreasing dimension of exit surface via changing the shape of aperture surface  $S_0$ . The aperture surface is changed from a flat to a Cartesian oval. This technique helps to decrease dimension of exit surface ( $S_2$ ). Consequently, the light beam can reach to receiver easier. The groove design process can be described by using Fig 12.

The reducing of  $S_2$  dimension can be described using the conservation of the optical path length and edge ray theorem. The light beam, which is refracted at  $S_0$  (Cartesian oval surface), focuses at position of the point  $A$ . Therefore, the  $S_0$  surface can be built by selection of position  $A$  and

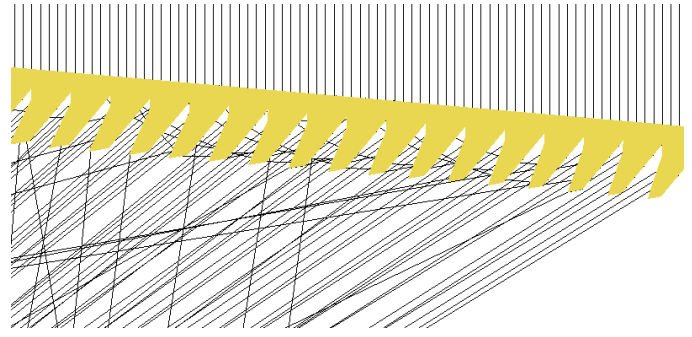


Fig. 11. The ray tracing of outer area is performed in Lighttools<sup>TM</sup> software and the optical loss in this part.

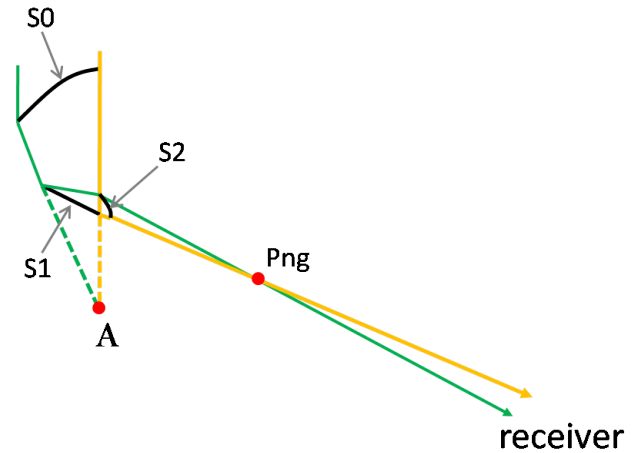


Fig. 12. The irradiance distribution of groove, which has  $S_0$  as Cartesian oval surface.

using equation of optical path length. In fact, the light beam reflects at TIR surface  $S_1$  before arriving position  $A$ . Thus, the irradiance distribution of light beam can be described gradually using the Edge ray theorem again. In this design, the reflection angle at  $S_1$  decreases gradually from the left edge ray to the right edge ray. Hence, the dimension of exit surface  $S_2$  can be reduced. In addition, it is similar to the middle area, the position of the focal point  $P_{ng}$  plays an important role to distribute uniformly over the receiver. The  $P_{ng}$  can be estimated using conservation of optical path length. After design process, the shape of outer area of the designed lens is shown in Fig. 13.

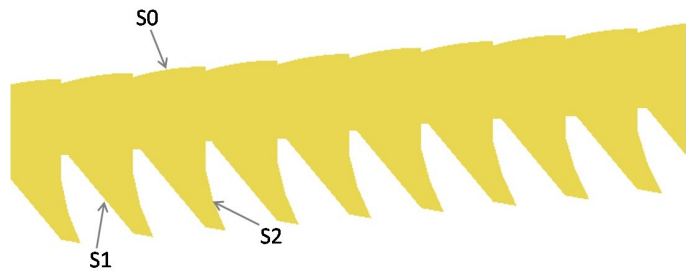


Fig. 13. The optimum shape of the outer area.

The estimation of the outer area's limitation is similar to the middle area. The  $S_2$  surface tends to move to  $S_1$  surface when diameter of lens increases. As a result, the limitation of the outer area can be estimated when  $S_2$  intersect  $S_1$  surface. When the position of the  $S_2$  intersects to  $S_1$  surface, the size of the designed lens is estimated completely.

After all, the design method of a new structured Fresnel lens is introduced. In addition, a new technique is also proposed to achieve uniform irradiance distribution over receiver.

### III. PERFORMANCE AND DISCUSSION

In this design, Matlab constructs the shape of designed Fresnel lens. Then the optimum structure of the designed Fresnel lens is estimated using rays tracing process in LightTools<sup>TM</sup> software. Furthermore, in order to examine the new design method, the simulation process is also performed using LightTools<sup>TM</sup>. The Table I shows some light source parameters of ray tracing and simulation.

TABLE I  
RAYS TRACING AND SIMULATION PROCESS PARAMETERS.

Items	Values
Wavelength for rays tracing	550nm
Spectrum of light source	380nm – 1600nm
Power of light source	5000W
Material of designed Fresnel lens	PMMA
Refractive index	1.492

Fig. 14 shows that the designed Fresnel lens shape is drawn in 3-D consisting of three parts: inner, middle, and outer areas. The inner area is constructed by the smallest grooves, which is designed using refraction phenomena. Meanwhile, the middle area designed byb using TIR includes some bigger grooves. Finally, the outer area, which consists of the biggest grooves, is also built using TIR. The designed Fresnel lens is designed with wide grooves to show the profile clearly.

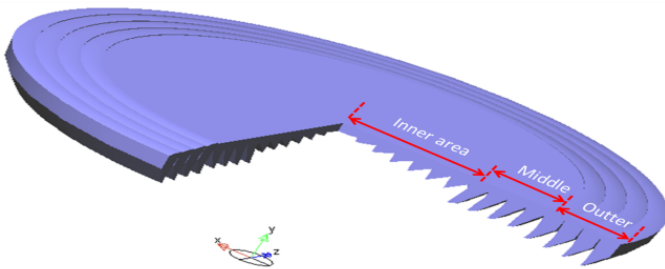


Fig. 14. The designed Fresnel lens shape in 3-D in LightTools<sup>TM</sup>.

Fig. 15 shows the shape of grooves in three parts of designed Fresnel lens.

- The inner groove has exit surface as a Cartesian oval. The light beam is redirected by refracting at exit surface to distribute uniformly over receiver.
- Meanwhile, the light beam reflects at flat surface  $S_1$  after that it refracts at exit surface  $S_2$  to go to receiver in the middle area.

- The irradiance distribution in the outer area is more complicated than that in the inner and the middle area. The light beam is refracts two times at aperture, exit surface and reflects at TIR surface.

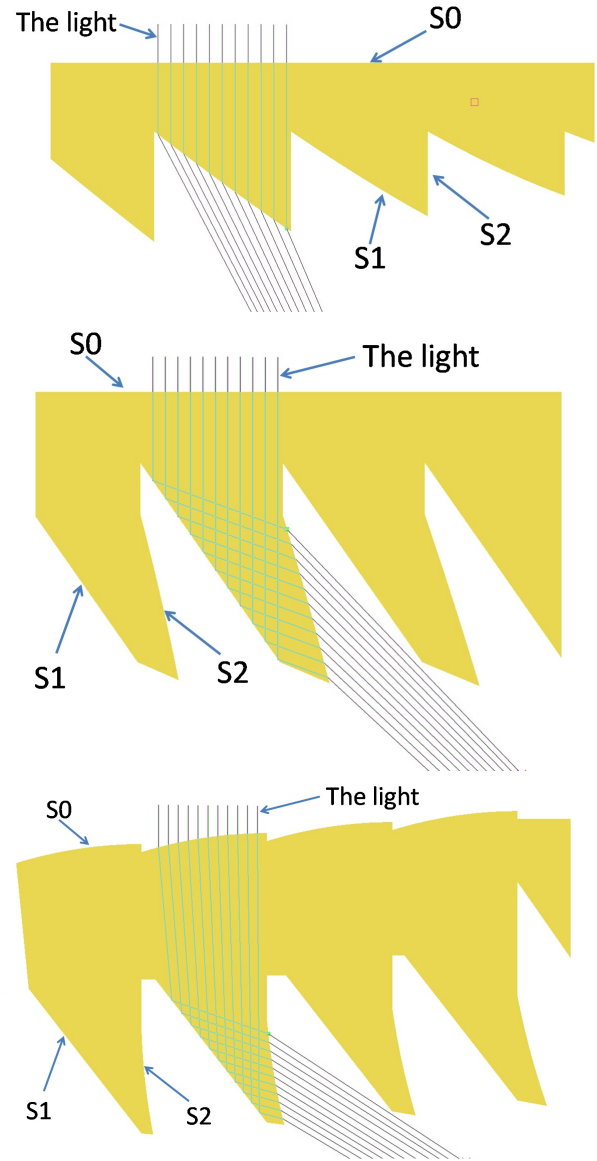


Fig. 15. The shape of a) the conventional groove and b) cutting groove to reduce optical loss.

The ray tracing results of designed Fresnel lens are demonstrated in Fig. 16. The direct sunlight, which comes to the designed Fresnel lens, is collected and distributed uniformly over the receiver. In addition, the boundaries among three parts are also shown in Fig. 16 as follows.

In order to observe the effect of the new design methods, the efficiency investigation is performed with some parameters of designed Fresnel lens in the Table II. In this technique, the concentration ratio is increased by extending diameter of Fresnel lens while the height of lens is a constant. The focal length of lens 140 mm is chosen to similar with some



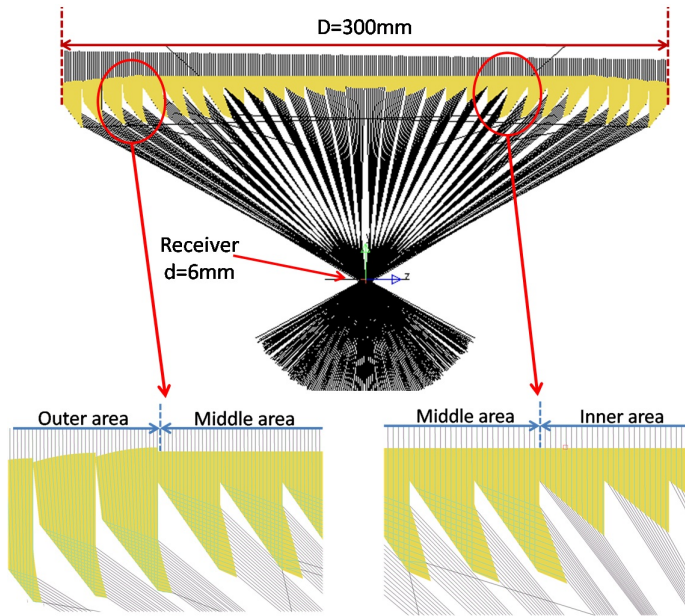


Fig. 16. The ray tracing of designed Fresnel lens.

conventional Fresnel lens (usually from 100 mm to 400 mm) [13], [20], [21]. The conventional Fresnel lens normally has f-number  $\approx 1$  so that the lens with focal length 140 mm also has diameter equal 140 mm. Therefore, the diameter of the proposed Fresnel lens is chosen 140 mm in beginning then it is increased with 220 mm, 300 mm, and 400 mm to increase concentration ratio.

TABLE II  
DESIGN PARAMETERS FOR FRESNEL LENS.

Items	Values
Dimension of designed Fresnel lens	140, 220, 300, 400 mm
Width of groove	2 mm
Height of lens	140 mm
Geometrical concentration ratio	545x, 1345x, 2500x, 4444x
Number of grooves	70, 110, 150, 200
F-number	1, 0.63, 0.46, 0.35
Diameter of receiver (solar cell)	6 mm [21]

The optical efficiency is an important parameter in CPV and it is defined by the rate between irradiance coming to Fresnel lens and the irradiance reaching to receiver [6]. In Fig. 17, the relationship between optical efficiency and f-number of designed Fresnel lens is illustrated using LightTools<sup>TM</sup> when the concentration ratio of lens increases. The results demonstrate that the designed Fresnel lens is a good choice to collect sunlight in CPV system because of high concentration ratio, high optical efficiency ( $\approx 86\%$  at 2500x), and small f-number (0.46).

Beside concentration ratio and optical efficiency, the uniformity of irradiance distribution is also an important parameter in CPV system. Therefore, the irradiance distribution is investigated in design process. Fig. 18 shows the uniform irradiance distribution of designed Fresnel lens over the receiver.

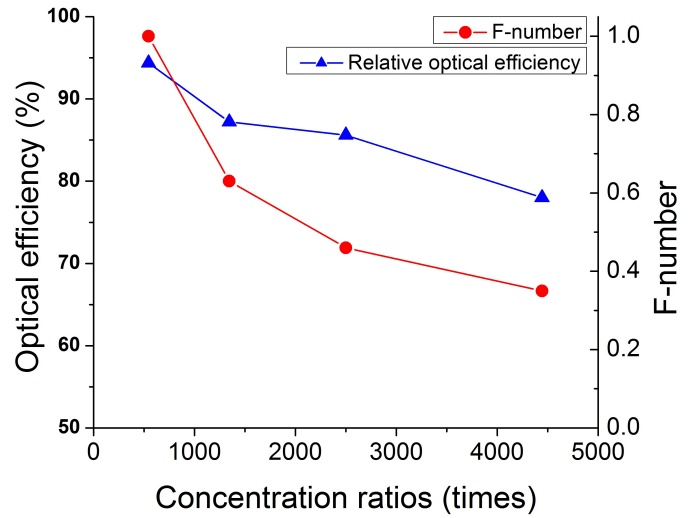


Fig. 17. The optical efficiency and f-number at different concentration ratios.

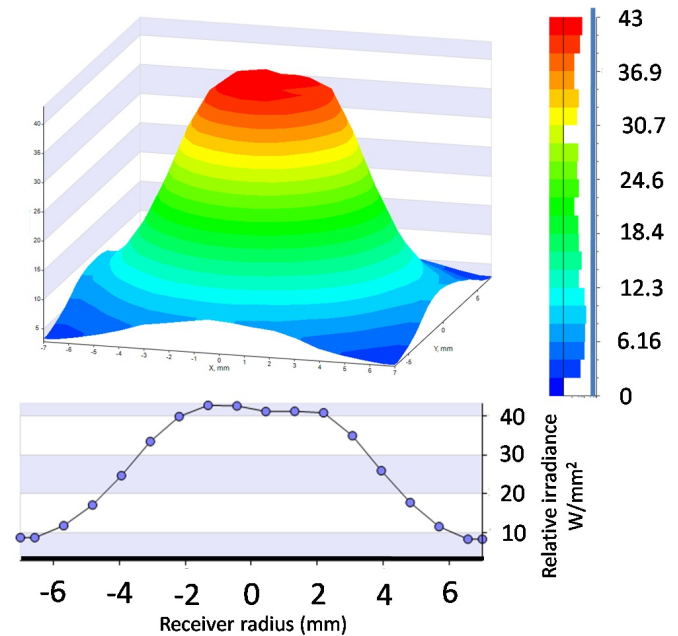


Fig. 18. The uniform distribution of irradiance over the receiver in designed Fresnel lens.

The irradiance distribution of the designed Fresnel lens is uniform so that there is no any hot spot point over the receiver. However, there is a small spreading of distribution spectrum at the bottom (see Fig. 18). That can be explained by the light source, which has a wide range of wavelength from 380 nm to 1600 nm. Thus, the active area of receiver is bigger than the size of ideal receiver (or designed receiver) in real. Hence, the real concentration ratio of lens always reduces few times. In this case, the ratio concentration of lens is decreased from 2500x to 900x for lens with diameter  $D = 300$  mm because of increasing of receiver diameter from  $d = 6$  mm to  $d = 10$  mm. Nevertheless, the lens with concentration ratios 900x is still a



perfect choice at the recent technology of multi-junction solar cells [22].

In the CPV, the acceptance angle is defined as the incident angle at which the solar power over receiver drops to 90% of its maximum [23], [24]. Besides, the acceptance angle of the concentrator must be larger than the solar angle  $0.275^\circ$  [9]. Thus, the acceptance angle should be measured in this design. The result in the simulation shows that the acceptance angle of CPV using designed Fresnel lens is  $0.7^\circ$  without SOE. Recently, the advantages of tracking technology help to improve the accuracy of CPV system [25]. Therefore, the acceptance angle  $0.7^\circ$  without SOE is a good candidate to apply to concentrator photovoltaic.

#### IV. CONCLUSION

The study proposed new technique using the conservation of optical path length and the edge ray theorem to design Fresnel lens, which consists of three parts: the inner area, the middle area, and the outer area. This structured lens can achieve good properties such as optical efficiency 86% at high concentration ratio (900x), high uniform irradiance distribution, small f-number (0.46), and effective acceptance angle ( $0.7^\circ$ ) without SOE. In addition, the simulation was performed using the light source with the wide range of wavelength from 380nm – 1600nm to mimic the real conditions, at which the designed Fresnel lens have to work. Furthermore, the way of design process using refraction and reflection phenomena was described in details. All these factors substantially help to enhance the performance of CPV and decrease its related cost when mass installing. The concentrator based on the designed Fresnel lens is a promising option for development of the high performance and cost-effective concentrator photovoltaic system generation. Therefore, the near future task is establishment of CPV system using designed Fresnel and it will be compared with different structures, which exist in the market.

#### REFERENCES

- [1] Fraunhofer, I. S. E., "New World Record for Solar Cell Efficiency at 46% French-German Cooperation Confirms Competitive Advantage of European Photovoltaic Industry," (2014).
- [2] Mehrdad Khamooshi, Hana Salati, Fuat Egelioglu, Ali Hooshyar Faghiri, Judy Tarabishi, and Saeed Babadi, "A Review of Solar Photovoltaic Concentrators," *International Journal of Photoenergy* 2014, 958521 (2014).
- [3] D. Abbott, "Keeping the energy debate clean: how do we supply the world's energy needs?" *Proc. IEEE* 98, 42–66 (2010).
- [4] B. Mendoza, "Total solar irradiance and climate," *Advances in Space Research* 35, 882–890 (2005).
- [5] E. Lorenzo and A. Luque, "Comparison of Fresnel lenses and parabolic mirrors as solar energy concentrators," *Applied Optics* 21, 1851–1853 (1982).
- [6] Ralf Leutz, Akio Suzuki, Atsushi Akisawa, Takao Kashiwagi "Design of A Nonimaging Fresnel Lens for Solar Concentrators," *Solar Energy* 65, 379–387 (1999).
- [7] Juan C. González, "Design and analysis of a curved cylindrical Fresnel lens that produces high irradiance uniformity on the solar cell," *Applied Optics* 48, pp. 2127–2132 (2009).
- [8] N. Yamada and T. Nishikawa, "Evolutionary algorithm for optimization of nonimaging Fresnel lens geometry," *Optics Express* 18, A126–A132 (2010).
- [9] Jui-Wen Pan, Jiun-Yang Huang, Chih-Ming Wang, Hwen-Fen Hong, Yi-Ping Liang, "High concentration and homogenized Fresnel lens without secondary optics element," *Optics Communications* 284, 4284–4288 (2011).
- [10] P. Espinet, C. Algora, I. Rey-Stolle, I. Garcia and , M. Baudrit, "Electroluminescence Characterization of III-V Multi-junction Solar Cells," *Proc. IEEE PVSC 33rd* 4922477, 1–6 (2008).
- [11] Lei Jing, Hua Liu, Hui-fu Zhao, Zhenwu Lu, Hongsheng Wu, He Wang, and Jialin Xu, "Design of novel compound Fresnel lens for High-performance photovoltaic concentrator," *International Journal of Photoenergy* 2012, 630692 (2011).
- [12] Stanislas Sanfo, Abdoulaye Ouedraogo, "Contribution to the Optical Design of A Concentrator with Uniform Flux for Photovoltaic Panels," *Advances in Energy and Power* 3, 82–89 (2015).
- [13] Irfan Ullah and Seoyong Shin, "Development of Optical Fiber-based Daylighting System with Uniform Illumination," *Journal of the Optical Society of Korea* 16, 247–255 (2012).
- [14] Kwangsun Ryu, Jin-Geun Rhee, Kang-Min Park, Jeong Kim, "Concept and design of modular Fresnel lenses for concentration solar PV system," *Solar energy* 80, 1580–1587 (2006).
- [15] Atsushi Akisawa, Masao Hiramatsu, Kouki Ozaki, "Design of dome-shaped non-imaging Fresnel lenses taking chromatic aberration into account," *Solar Energy* 86, 877–885 (2012).
- [16] Fernando Eismann, "Design of a plastic aspheric Fresnel lens with a spherical shape," *Optical engineering* 36, 988–991 (1997).
- [17] Roland Winston, Juan C. Miñano and Pablo Benítez, *Non-imaging optics* (Elsevier Academic Press, Burlington, MA 01803, USA, 2005), Appendix B.
- [18] Leutz and Suzuki, *Non-imaging Fresnel lens, Design and performance of solar concentrators* (Springer series in optical sciences, Berlin Heidelberg, Germany, 2001).
- [19] Katsuki Tanabe, "A review of ultrahigh efficiency III-V semiconductor compound Solar cells: Multijunction Tandem, lower dimensional, photonic Up/Down conversion and plasmonic nanometallic structures," *Energies* 2, 504–530 (2009).
- [20] Kok-Keong Chong, Sing-Liong Lau, Tiong-Keat Yew, Philip Chee-Lee Tan, "Design and development in optics of concentrator photovoltaic system," *Renewable and Sustainable Energy Reviews* 19, 598–612 (2013).
- [21] W. T. Xie, Y. J. Dai, R. Z. Wang, K. Sumathy, "Concentrated solar energy applications using Fresnel lenses: A review," *Renewable and Sustainable Energy Reviews* 15, 2588–2606 (2011).
- [22] Pablo Benítez, Juan C. Minano, Pablo Zamora, Ruben Mohedano, Aleksandra Cvetkovic, Marina Buljan, Julio Chaves, Maikel Kernandez, "High performance Fresnel-based photovoltaic concentrator," *Optical Society of America* 18, S1/Optics Express A25 (2010).
- [23] Marina Buljan, João Mendes-Lopes, Pablo Benítez, Juan Carlos Miñano, "Recent trends in concentrated photovoltaics concentrators' architecture" *Journal of Photonics for Energy* 4, 040995-1 (2014).
- [24] A. Yavrian, S. Tremblay, M. Levesque and R. Gilbert, "How to increase the efficiency of a high concentrating PC (HCPV) by increasing the acceptance angle to  $\pm 3.20^\circ$ ," *Proc. AIP Conference Proceedings* 1556, 197–200 (2013).
- [25] Minor M. Arturo, Garcia P. Alejandro, "High-Precision Solar Tracking System" in *Proc. The World Congress on Engineering* (London, UK, July. 2010) pp. 844–846.

# Skin Lesion Segmentation based on Integrating EfficientNet and Residual block into U-Net Neural Network

Duy Khang Nguyen, Thi-Thao Tran, Cong Phuong Nguyen, Van-Truong Pham\*

*School of Electrical Engineering  
Hanoi University of Science and Technology  
Hanoi, Vietnam  
truong.phamvan@hust.edu.vn*

**Abstract**—Skin lesion segmentation is an important step in computer aided diagnosis for automated melanoma diagnosis. However, in the field of medical images analysis, skin lesion segmentation from dermoscopic images is still a challenging task because of presence of various artifacts, blurring and irregular edges of the lesion. This paper proposes an efficient deep learning-based approach for skin lesion segmentation. Particularly, the paper proposes an improved version of the U-Net to perform skin lesion segmentation tasks. To this end, we propose to utilize EfficientNetB4 in encoder part of the original U-Net. In addition, the decoder part of the proposed network is constructed by residual block from Resnet architecture. By this way, the proposed approach could take advantages of the EfficientNet and Resnet architectures such as preserving efficient reception field size for the model, and avoiding the overfitting problem. The proposed approach is applied to segment images from ISIC 2017 and 2018 datasets. Experimental results show the desired performances of the proposed approach in terms of metrics of Dice coefficient and Jaccard indexes.

**Keywords**— Skin lesion segmentation, Deep learning, Image segmentation, Skin cancer, Deep neural networks

## I. INTRODUCTION

Skin cancer is one of the most widespread cancer types in over the world [1]. There are different types of skin cancer such as basal cell carcinoma, melanoma, intraepithelial carcinoma, squamous cell carcinoma, etc. [2]. Among them, malignant melanoma is a common and threatening skin cancer. It also is one of the most rapidly increasing cancers all over the world. However, determining the kind of lesions with the naked eye is a challenging task because the benign and malignant skin lesions are visual similarities. Over the years, there has been a rising interest in Dermoscopy—a non-invasive imaging technique that allows us to visualize skin surfaces by the light magnifying device and immersion fluid [3]. Nevertheless, the sole use of human vision to detect melanoma in dermoscopic images might be inaccurate, subjective, and irreproducible since it depends on the dermatologist's experiences [4].

To handle the above difficulties encountered in the diagnosis of melanoma, computer-aided diagnosis (CAD) systems are developed to assist the experts in the diagnosis process. There are four steps in CAD systems for identifying a lesion as melanoma: preprocessing, segmentation, feature extraction, and classification, in which lesion segmentation is a fundamental but very important step [5]. However, lesion segmentation remained challenging tasks due to the large variety in size, shape, and color along with different types of

skin and texture [4,5]. In addition, some lesions have irregular and fuzzy borders, and in some cases, the contrast between lesion and the surrounding skin is low. Moreover, artifacts and intrinsic cutaneous features, such as hairs, blood vessels and air bubbles might make the automatic segmentation more challenging.

Automated skin lesion segmentation research goes as early as the late 1980s. In the early days, research was mostly focused on hand-built features using statistics or math algorithm, such as transforming RGB colors into spherical color space using coordinates, transforming principal components in a user selected color space [6], or by finding first the average color of a small area of a lesion [7]. Though hand-built features made some progress in lesion segmentation, this process was time-intensive and not efficient for larger datasets. To tackle the automated skin lesion segmentation problem, there are some efforts proposed in the literature. Celebi et al. [8] proposed to use ensembling four different thresholding methods to detect the skin lesion border. Peruch et al. [9] applied color histogram and a thresholding technique to cluster the pixels in interested images into lesional and non-lesional skin. Zhou et al. [10] employed the improvement snake model to deal with skin lesion segmentation.

Recently, deep learning-based methods especially convolutional neural networks (CNNs) have obtained significant success in both image classification and segmentation task. CNNs are supervised training models which are trained to learn hierarchies of features automatically [11]. CNNs have been successfully applied to solve different image segmentation problems. Later than, improvement architectures of CNNs such as Alex-Net [12], VGG-Net [13], Google-Net [14] and Res-Net [15], have made the CNN to be standard for classification in many applications. Another CNN designed for biomedical image segmentation is U-Net [16]. Architecture of U-Net is similar to a U-shape and consists of a contracting path to capture context and a symmetric expanding path that enables precise localization of biomedical objects [16].

In this study, based on U-Net architecture, we propose an improved version of the U-Net and applied it to perform skin lesion segmentation tasks. To this end, we propose to utilize EfficientNetB4 in encoder part of the original U-Net. In addition, the decoder part of the proposed network is constructed by residual block from Resnet architecture. By this way, the proposed approach could take advantages of the EfficientNet and Resnet architectures such as preserving efficient reception field size for the model, and avoiding

overfitting and the gradient vanishing problem. The proposed approach is applied to segment images from ISIC 2017 and 2018 datasets.

## II. MATERIALS AND METHODS

### A. Datasets

This study is evaluated on two public datasets, the ISIC 2017 and the 2018 datasets [17]. The training data in 2017 dataset includes 2000 RGB dermoscopic images in various resolution and corresponding grayscale masks. The masks indicate the label for every pixel in dermoscopic images that the pixels belong to lesion areas or not. The test dataset consists of 600 similar pairs used for final evaluation. In the training phase, we shuffled the train dataset and split it in 80-20 ratio, the use of the small piece, to be validation dataset. All of images and masks are resized to 256x256 resolution.

The other dataset for lesion segmentation-ISIC 2018 provides 2594 images with ground truths for training [18,19]. There are about 100 images for validation and 1000 images for testing, but the ground truths for validation and testing images are not provided. Therefore, we divided the training data into two parts with ratio 4:1 as training sets and validation sets. We then also resize each image (for training and validation) to the 256x256 pixels. For training set, during training we use some standard data augmentation including random rotations and flips. Sample images are shown in **Fig.1**.

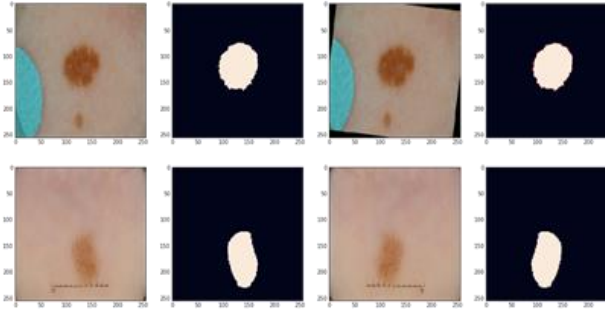


Fig. 1. Some representative images and ground truths with augmentation from the dataset

### B. Network architecture

A U-Net neural network segmentation architecture was shown in **Fig. 2**. In general, a U-Net architecture contains two paths [16]. The first path is the contraction path (also called as the encoder), which is used to capture the context in the image. The second path is the symmetric expanding path (also called as the decoder), which is used to enable precise localization using transposed convolution. In order to localize and upsample features, the expansive path combines them with high-resolution features from the contracting path via skip-connections. The output of the model is a pixel-by-pixel mask, that show the predict class of output for each pixel. U-Net is an end-to-end fully convolutional network (FCN). U-Net architecture proved itself very useful for segmentation problems with limited amounts of data.

In this study, we proposed a modification version of the U-Net model. The architecture of the proposed model is shown in **Fig. 3**, in which we have made some improvements in encoder and decoder parts of original U-Net model. In particular, in encoder part, we employ EfficientNets [20] which is formed from mobile inverted bottleneck convolution layer (MBConv) with an expansion ratio of 1 and 6 respectively (MBConv1 and MBConv6). EfficientNets are a family network structure, achieved by scaling network in all three dimensions: width, depth and height of base-line network EfficientNet-Bx. If we increase the network by, we need to scale up width, depth and resolution in the following way:

$$d = \alpha^\phi, w = \beta^\phi, r = \gamma^\phi \quad (1)$$

$$\text{with } \alpha.\beta^2.\gamma^2 \approx 2 \text{ and } \alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

Example EfficientNet-B0 is baseline network, we have  $\phi=0, d=1, w=1, r=1$ . To obtain EfficientNet-B1, we chose  $\phi=1, d=\alpha, w=\beta, r=\gamma$ . These networks of family achieve dramatically better efficiency than previous ConvNets including ResNets, DenseNets, and Inception [20]. In this study, we scale up the baseline and use EfficientNet-B4 for backbone of U-Net structure for skin lesion segmentation problem.

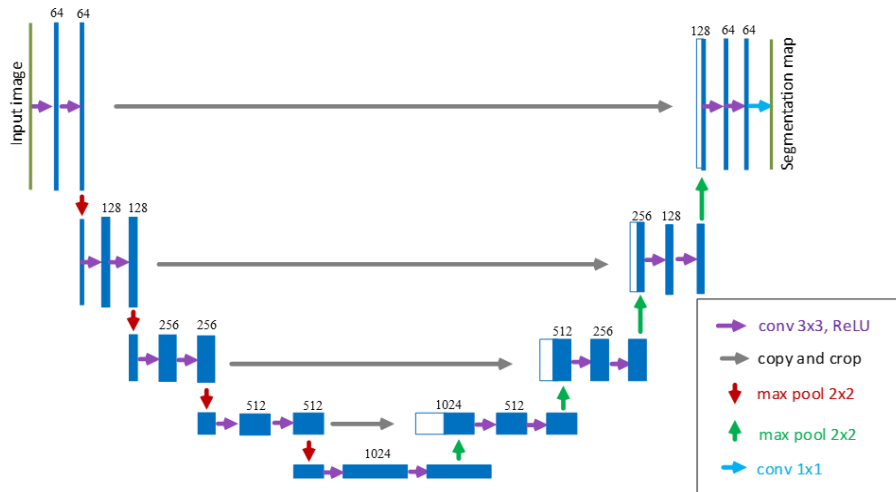


Fig. 2. U-Net architecture

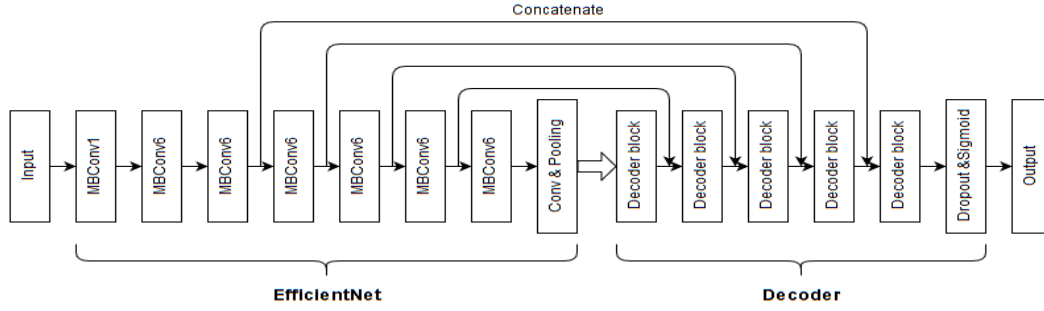


Fig. 3. The proposed network architecture

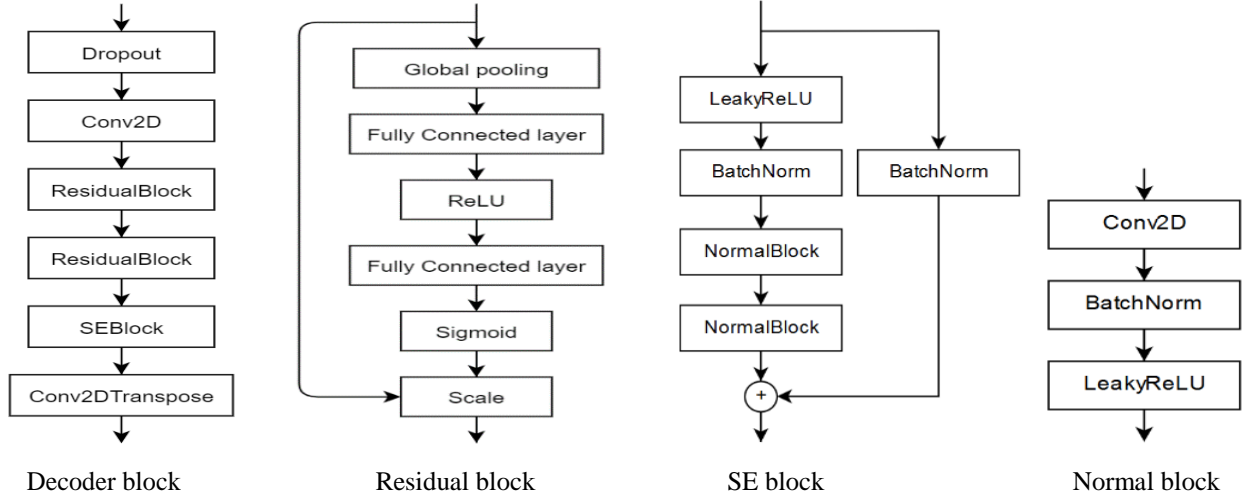


Fig. 4. Material block of decoder part of the proposed network

For the decoder part of the network, in this study, it is constructed by residual block from Resnet architecture to avoid overfitting and gradient vanishing. Thus, the proposed network has skip-connection in both encoder and decoder parts. Detail of the blocks of decoder part of the proposed network are presented in **Fig. 4**

### C. Loss function

As for loss function, we choose binary cross-entropy and dice loss. Let  $\hat{y}$  be the output of the last network layer passed through a sigmoid activation and  $y$  be the corresponding label. The binary cross-entropy (BCE) is then defined as follows

$$BCE = -\frac{1}{n} \sum_{i=1}^n (y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)) \quad (2)$$

and the dice loss is:

$$Dice = \frac{-2 \sum_i \hat{y}_i y_i}{\sum_i y_i + \sum_i \hat{y}_i} \quad (3)$$

Finally, the loss function is given as follows:

$$J = BCE + Dice \quad (4)$$

## III. RESULTS

### A. Evaluation

For quantitative analysis of segmentation result, several method were considered. These are Dice similarity

coefficient (DSC) and Jaccard (JAC) coefficient. The Dice coefficient is represented in the equation (5), where  $S_a$ ,  $S_m$  and  $S_{am}$  are, respectively, the automatically delineated region from datasets, the manually segmented region from our network, and the intersection between the two regions:

$$DSC = \frac{2S_{am}}{S_a + S_m} \quad (5)$$

In addition, Jaccard coefficient that is calculated according to the following equation, also used to measure dissimilarity between two sets:

$$JAC = \frac{S_{am}}{S_a + S_m - S_{am}} \quad (6)$$

### B. Results

We first evaluate the segmentation performances of the proposed approach on the ISIC 2017 database. To this end, we applied the proposed model to segment images from the ISIC 2017 dataset. Several samples obtained from ISIC2017 dataset were represented in the **Fig. 5**. In this figure, the original images, the segmentation by the proposed network and the ground truths of images are show from first to last rows in the order. From this figure it is observed that Combining EfficientNet and Residual block into U-Net provide good segmentation results.

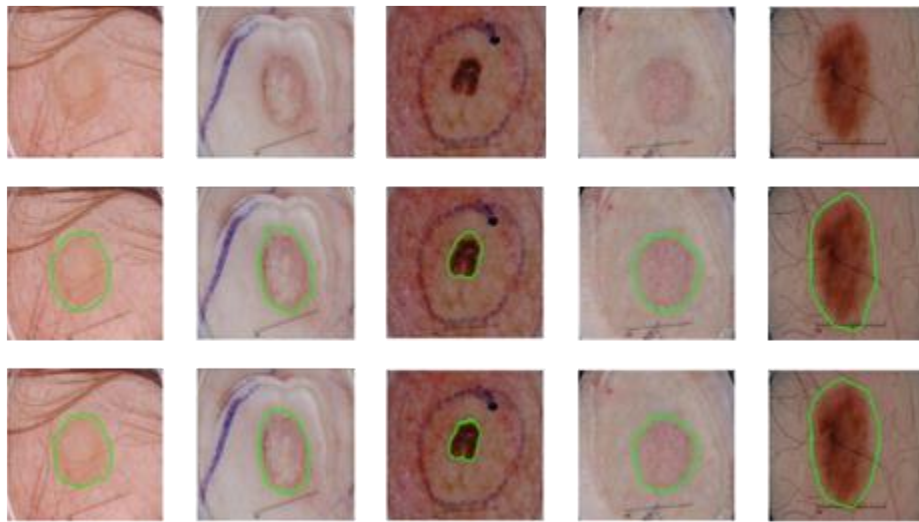


Fig. 5. Representative segmentation results on the ISIC 2017 database.

*Top: input images; middle: segmentation results by the proposed approach; bottom: ground truths*

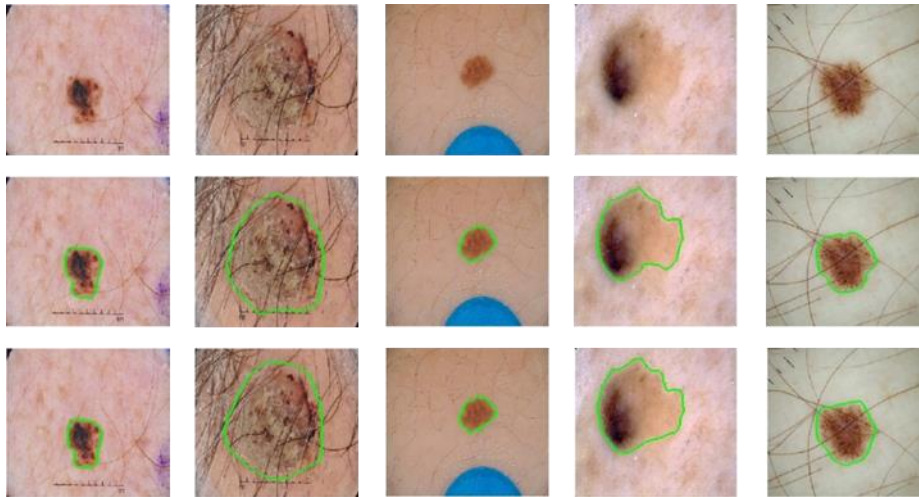


Fig. 6. Representative segmentation results on the ISIC 2018 database.

*Top: input images; middle: segmentation results by the proposed approach; bottom: ground truths*

**Table 1** summarizes the performance of our methods when compared with other state-of-the-art methods included Galdran [21], Jahanifar et al. [22], Bi et al. [23], Xue et al. [24], Ronneberger [16], FCN [11], Yuan et al. [25], and Ninh et al. [26] methods. We use the average Dice Similarity Coefficients, and Jaccard coefficients of each method when evaluate performance. We could see that the proposed method model achieved more accurate than other comparative methods.

In the next experiment, the proposed approach is used to segment images from ISIC2018 dataset. Some representative samples of segmented results for ISIC2018 dataset are presented in **Fig. 6**. The original images, the segmentation by the proposed method are shown in the first and second row of the figure, respectively, in this figure. The ground truths corresponding to images in the first row are also given in the last row. From this figure, we see that, there is a good agreement between the results obtained results and the ground truths, which is demonstrated the advantages of the proposed approach.

TABLE 1. THE MEAN OF OBTAINED DICE SIMILARITY COEFFICIENT (DSC) AND JACCARD COEFFICIENT (JAC) BETWEEN OTHER STATE-OF-THE-ART AND THE PROPOSED MODELS ON THE ISIC CHALLENGE 2017 DATASET

Method	Dice Coefficient	Jaccard Coefficient
Galdran method [21]	0.810	0.718
Jahanifar et al. method [22] (2019)	0.827	0.721
Bi et al. method [23]	0.834	0.731
Xeu et al. method [24]	0.839	0.749
Ronneberger et al. method [16]	0.842	0.758
FCN method [11]	0.827	0.721
Yuan et al. method [25] (2019)	0.849	0.765
The method [26] (2020)	0.853	0.771
The proposed approach	<b>0.861</b>	<b>0.781</b>



In the **Table 2**, we present the performance of our method with two other model: Original UNet and MaskR-CNN– 1<sup>st</sup> place networks in ISIC2018 challenge. We evaluate and analyse performance of our algorithm using Dice Coefficients, and Jaccard coefficients and specify computation cost, including number of parameter and training time of these networks.

TABLE 2. THE MEAN OF OBTAINED DICE SIMILARITY COEFFICIENT (DSC) AND JACCARD COEFFICIENT (JAC) ON THE ISIC CHALLENGE 2018 DATASET BY THE PROPOSED MODEL AND RECENT WORKS

Network structure	Dice Index	Jaccard Index
Original UNet [16]	0.803	0.720
MaskR-CNN [27]	0.898	0.838
The proposed approach	<b>0.912</b>	<b>0.846</b>

It can be observed from the Table 2, the proposed achieved highest scores in the default performance measures in this challenge when compared to the other algorithms.

#### IV. CONCLUSION

We have presented a modification of U-Net-based model for skin lesion segmentation. Particularly, the proposed approach utilizes EfficientNetB4 in encoder part of the original U-Net. Meanwhile, the decoder part of the proposed model is constructed by residual block from Resnet architecture. The proposed approach therefore can preserve efficient reception field size for the model, and avoid the overfitting problem. The proposed approach reveals the desired performances when applied to segment images from ISIC 2017 and 2018 datasets.

#### ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2018.302.

#### REFERENCES

- [1] Pathan, S., Prabhu, K.G., Siddalingaswamy, P.: Techniques and algorithms for computer aided diagnosis of pigmented skin lesions: A review. *Biomedical Signal Processing and Control* **39**, 237-262 (2018).
- [2] Gandhi, S.A., Kampp, J.: Skin Cancer Epidemiology, Detection, and Management. *Med Clin. N. Am.* **99**(6), 1323–1335 (2015).
- [3] Pellacani, G., Seidenari, S.: Comparison between morphological parameters in pigmented skin lesion images acquired by means of epiluminescence surface microscopy and polarized-light videomicroscopy. *Clin Dermatol.* **20**(3), 222-227 (2002).
- [4] Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D.: Dermoscopic image segmentation via multi-stage fully convolutional networks. *IEEE Trans Biomed Eng.* **64**(9), 2065-2074 (2017).
- [5] Ünver, H., Ayan, E.: Skin Lesion Segmentation in Dermoscopic Images with Combination of YOLO and GrabCut Algorithm. *Diagnostics (Basel)* **9**(3), E72 (2019). doi:10.3390/diagnostics9030072
- [6] Kopf, A., Salopek, T., Slade, J., Marghoob, A., Bart, R.: Techniques of cutaneous examination for the detection of skin cancer. *Cancer* **75**(2), 684-690 (1995).
- [7] Umbaugh, S., Moss, R., Stoecker, W.: Automatic color segmentation of images with application to detection of variegated coloring in skin tumors. *IEEE Eng Med Biol Mag.* **8**(4), 43-50 (1989).
- [8] Celebi, M., Wen, Q., Hwang, S., Iyatomi, H., Schaefer, G.: Lesion border detection in dermoscopy images using ensembles of thresholding methods. *Skin Res. Technol.* **19**(1), e252–e258 (2013).
- [9] Peruch, F., Bogo, F., Bonazza, M., Cappelleri, V., Peserico, E.: Simpler, faster, more accurate melanocytic lesion segmentation through MEDS. *IEEE Trans Biomed Eng.* **61**(2) (2014).
- [10] Zhou, H., Li, X., Schaefer, G., Celebi, M.E., Miller, P.: Mean shift based gradient vector flow for image segmentation. *Comput. Vis. Image Understand.* **117**(9), 1004–1016 (2013).
- [11] J. Long, E. Shelhamer, T. Darrell: Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440 (2015).
- [12] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Proc. Advances in neural information processing systems 2012*, pp. 1097–1105
- [13] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv Prepr. arXiv1409.1556*, (2014).
- [14] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: 2015, pp. 448–456
- [15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 2016*, pp. 770–778
- [16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* 2015, pp. 234-241
- [17] Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: *Proc. IEEE 15th Int. Symp. Biomed. Imag.* 2018, pp. 168–172
- [18] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). <https://arxiv.org/abs/1902.03368> (2019).
- [19] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 doi:10.1038/sdata.2018.161. (2018).
- [20] Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: *in Proc. 36th Int. Conf. Mach. Learn* 2019, pp. 6105–6114
- [21] Galdran, A., Alvarez-Gila, A., Meyer, M.L., Saratxaga, C.L., Araújo, T., Garrote, E., Aresta, G., Costa, P., Mendonça, A.M., Campilho, A.C.: Data-driven color augmentation techniques for deep skin image analysis. in *arXiv:1703.03702*, <https://arxiv.org/abs/1703.03702> (2017).
- [22] Jahanifar, M., Tajeddin, N.Z., Asl, B.M., Gooya, A.: Supervised saliency map driven segmentation of lesions in dermoscopic images. *IEEE J. Biomed. Health Inform.* **23**(2), 509-518 (2019).
- [23] Bi, L., Kim, J., Ahn, E., Feng, D.: Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. in *arXiv:1703.04197*, Available: <https://arxiv.org/abs/1703.04197> (2017).

- [24] Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X.: SegAN: Adversarial network with multi-scale L 1 loss for medical image segmentation. *Neuroinformatic* **16**(383-392) (2018).
- [25] Y. Yuan, Y.-C. Lo: Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. *IEEE J. Biomed. Health Inform.* **23**(2), 519-526 (2019).
- [26] Ninh, Q.C., Tran, T.T., Tran, T.T., Tran, T.A.X., Pham, V.T.: Skin Lesion Segmentation Based on Modification of SegNet Neural Networks. In: *Proc. 2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, Hanoi 2020, pp. 575-578
- [27] He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV)* 2017, pp. 2980–2988

# An Educational Transformative Sustainability Model Based On Modern Educational Technology

Xuan Thanh Pham<sup>1,2</sup>

<sup>1</sup>Faculty of Educational Management  
Dong Nai University  
Bien Hoa, Vietnam

<sup>2</sup>Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thanh@eco2.vn

ORID: 0000-0002-3187-5227

Anh Tho Mai

Faculty of Information Technology  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thoma@hcmute.edu.vn  
ORCID: 0000-0001-9159-2379

Anh Tuan Ngo

Institute of Online Education (UTEx)  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tuankti@hcmute.edu.vn  
ORCID: 0000-0002-2264-4187

**Abstract**— It is noticeable that most traditional schools in Vietnam, with a closed educational viewpoint, have not been able to keep up with the constantly updated requirements of education during the fourth industrial revolution. So, what is the solution to transform a school into a new technology-based open one, aiming to create an education ecosystem in the future? This article proposes an educational transformative sustainability model based on modern educational technology with the appropriate pathway according to the reality of education in Vietnam, with a view to enhancing education efficiency in the new context. By surveying the current status and applying the proposed model to the case study of Ho Chi Minh City University of Technology and Education (HCMUTE), the result shows that this pathway is easy to apply, stimulating the expansion of this research to other schools, aiming forward to building unique education ecosystems, which would create a characteristic education ecosystem of Vietnam when combined.

**Keywords**— Education Ecosystem, Modern Educational Technology, New Technology-based School, An Educational Transformative Sustainability Model.

## I. INTRODUCTION

*“Education is not preparation for life; education is life itself” (John Dewey).*

Such a saying proves why education is considered a means to achieve all sustainable development goals [1], and why it is specified in Goal 4: “Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.” This has changed education towards the development of the 21st-century generation (in terms of individuals) with the adaptability and proactivity [2] to lead the sustainable development progress during rapid societal changes [3].

In order to achieve sustainability education goals, schools, where educational solutions are implemented, are to change in all aspects, from their curriculum, educational activities and organizational culture to research and community relationships [4], especially in the context of globalization and the information and communication technology (ICT) revolution.

Educational transformation can be seen through the diversification of learning forms in schools, from traditional classrooms to computer-aided ones, to online learning (e-learning, m-learning) or blended learning... gradually moving forward to the open education trend in the form of massive open online courses (MOOC), along with learning systems with personal learning environment (PLE), creating a future education ecosystem as it shows the close links between the learning components with each other, and with the external learning environment through the movement of connected knowledge and technology environment; and it also shows personalization through establishing relationships in order to create an educational connection environment for personal development (Thanh, 2019) [5].

This is associated with the change in viewpoints on ICT in education, from e3-Learning (efficient, effective and engaging), which means technology enhances learning efficiency, to c3-Learning (collaborative, contextual, and connected learning) [6], which emphasizes individual learning with the ability to self-direct the learning process, proactivity, and self-responsibility, called heutagogy: self-determined learning [7, 8, 9], in the context of an intellectual, lifelong-learning society.

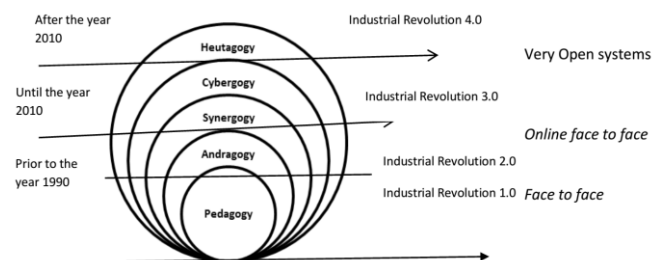
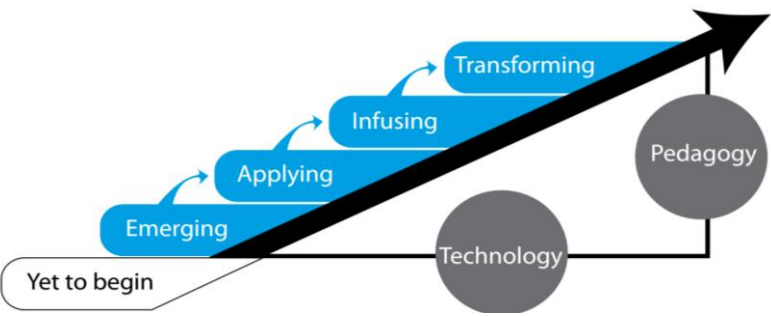
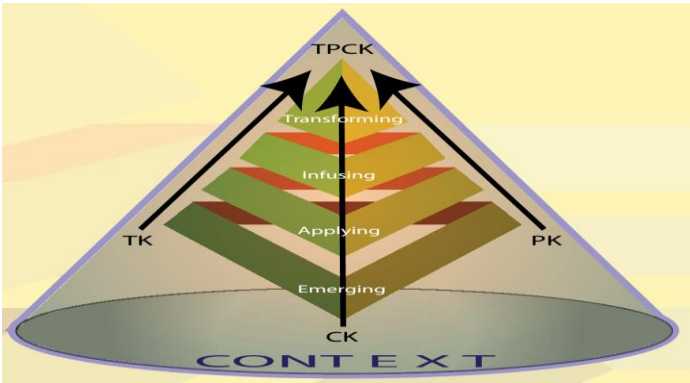
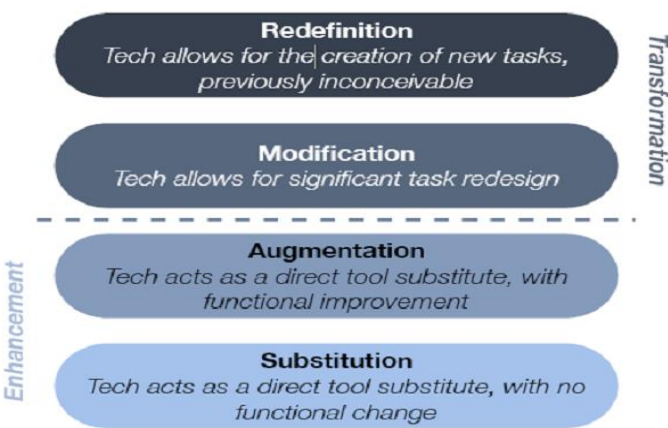


Fig. 1 Industrial Revolution and Education Revolution (source: [10])

The following table reflects the efforts to transform educational model in schools in the orientation to utilizing modern technology platforms. These models present the importance of transforming schools into new technology-based ones and that such transformation requires going through certain stages with an appropriate and practical pathway to enhance education efficiency in the new context.

TABLE 1. ICT IN EDUCATION MODELS FOR SCHOOL

Model's Figures	Descriptions
 <p>Fig. 2 Integration of ICT in education model (Anderson and van Weert, 2002; Anderson, Glenn, 2003; and Majumdar, 2005)</p>	<p>A model about integration of ICT in education which has two dimensions: technology and pedagogy. Within these two dimensions are seen four stages that classes or schools typically pass through in their integration of ICT [11].</p>
 <p>Fig 3. ICTeTD model (Engida, 2011)</p>	<p>ICT-enhanced teacher development model (ICTeTD) is technology use in teaching [12]. This one is a combination between 2 TPACK models, which are TPACK framework (Mishra and Koehler, 2006) and Transformation model (Anderson, 2002).</p>
 <p>Fig. 4 SARM model (R. R. Puetendura, 2010; Jude, L. T., Kajura, M. A., &amp; Birevu, M. P. , 2014)</p>	<p>SAMR model to assess ICT Pedagogical Adoption, describing an ICT led pedagogy to be the use of technology on different levels. These levels clearly depend on the user knowledge of integration and availability of the tools, categorized into 2 main stages, enhancement and transformation [13].</p>

There are other model development which also focuses on the technology factor, such as towards an educational model for lifelong learning [14], the innovative digital school model [15], and UNESCO ICT competency framework for teachers (UNESCO, 2011) [16].

In Vietnam, the transformation process into technology-based schools is methodically recorded by SEAMEO in the 2010 report [17] based on UNESCO's model of ICT development in education comprising of four stages and 10 ICT in education dimensions, which are the necessary and sufficient conditions that support the integration of ICT in education [18].

This report shows that Vietnam belongs to Group 2 - countries at the Infusing stage for most of the dimensions with ICT plans and policies in education. This acts an advantage, facilitating the first step to the transformation into new technology-based schools. However, the actual transformation taking place in Vietnam, in terms of educational technology, reflects the disconnection between technology and education, and the lack of dependence of technology on pedagogy basis. This results in low education efficiency and brings about significant barriers to a sustainable education ecosystem in the future [5].

Therefore, we believe that in order to successfully transform schools into new, sustainable technology-based ones, the following points should be carried out:

- Adjust UNESCO's model of ICT development in education to the reality of education in Vietnam, oriented towards modern educational technology.
- Propose a pathway for the transformation into new, sustainable technology-based schools.
- Apply the model to guide and adjust the transformation at HCMUTE (as a case study) in accordance with the pathway for sustainable development, creating an education ecosystem in the future.

## II. AN EDUCATIONAL TRANSFORMATIVE SUSTAINABILITY MODEL BASED ON MODERN EDUCATIONAL TECHNOLOGY

### A. The essence and necessity of modern educational technology to ensure sustainable development of technology-based schools

"Most authors underline educational technology as a systematic way, a process or an application of the scientific knowledge, to improve the efficiency of the process of learning and instruction. It is thus considered to be the technology of education more than the technology in education" (Kamar, 1996) [19].

The essence of educational technology presented by Kamar in the present context should be understood as (1) the

technology of education, manifested itself through the process of organizing teaching and learning activities in practice, which exists in all forms of education, reflecting the level of understanding and the capability to organize educational activities; and (2) the application of technological advancements into education (technology in education), diversifying forms of education, considered the technological component, alongside pedagogy, social, psychological, of modern education environment. Under the impact of the technological component, the inherent technological attribute of education must transform in accordance with the new context, and, thanks to such impact, the efficiency of technological attribute increases, leading to a remarkable increase in education efficiency.

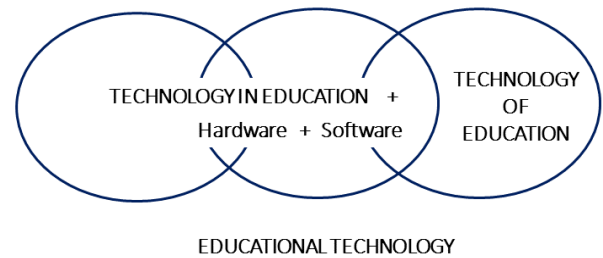


Fig. 5 Educational Technology (source [19])

This correlation reflects the characteristics of educational technology, summarized in the following table, demonstrating its impact on the development of new technology-based schools [5] [20].

TABLE 2. IMPACT OF EDUCATIONAL TECHNOLOGY ON THE TRANSFORMATIVE MODEL

Modern educational technology-based approach		Impact
Characteristics	Description, evidence	
<i>Integrity and synchronization</i>	Reflected in the correlation among major components of Educational Technology (4M: Method, Materials, ManPower & Media)	The solution should be comprehensive, synchronous, and consistent with the overall development goals of the school.
<i>Mutuality and the key factors to education effectiveness</i>	Reflected in the synchronous development and mutual support between hardware and software, in which software (instructional design) is the key factor to education effectiveness	The solution should focus on the development of teachers' competence in instructional technology. Ensure synchronous development between technological infrastructure and the capacity to use it efficiently.
<i>Openness, constant updates and flexibility</i>	Reflected in the constant updates of the technological component, requiring practical changes in education.	The solution requires a vision for the development of new technology-based schools (e.g. in the orientation to an education ecosystem)
<i>Diversity, creativity and differences without losing effectiveness</i>	Reflected in the appearance of many forms of teaching due to the diversity of technology.	The solution should ensure diversity, allowing schools to select the forms of teaching consistent with the reality and teachers' competence.
<i>Feasibility</i>	Reflected in the strategic implementation of instructional technology, including the following stages: analyze, design, implement, evaluate and adjust.	The solution should follow the instructional technology process to ensure success and optimal quality.



With this modern educational technology-based approach, Trung et al (2020) has proposed the eClass model, considered as the core of new technology-based schools on the basis of an education ecosystem (Fig 6). This model has been applied into practice at Dong Nai University in Pedagogy 2 course. Experimental results show that the model is effective in teaching. For example, it has helped transform the teaching model to develop vocational competency, change the evaluation methods into competency-based, and develop ICT and other competencies for students [5].

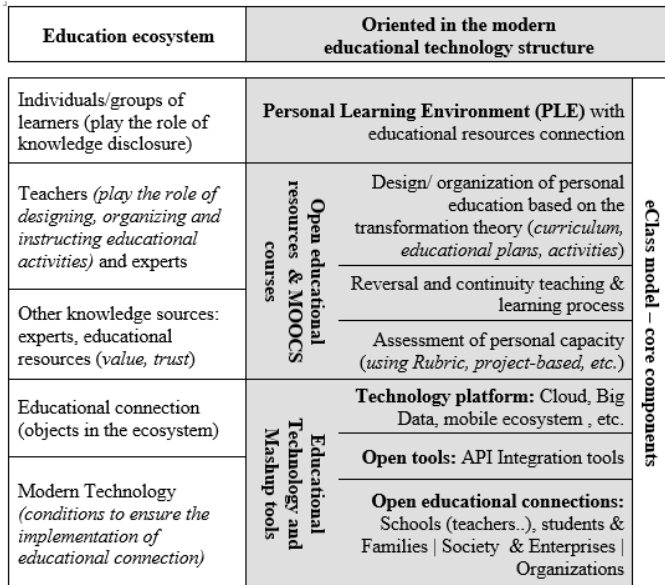


Fig 6. eClass model with core components (source [5])

This proves the feasibility and the necessity of modern educational technology-based approach to successfully and sustainably develop new technology-based schools, and is the destination of the transformative model hereinafter.

### B. The transformative model, adjusted to the modern educational technology-based approach

Following SEAMEO's report on ICT in education in Vietnam, in 2012, Vietnam conducted a 2nd evaluation to prepare for the development objectives in 2020 [17, 21]. The evaluation result is illustrated in Fig 7.

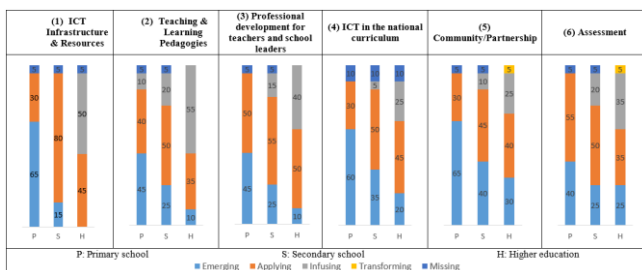


Fig 7. Evaluation of current status of ICT in education in primary, secondary and higher education in Vietnam (with first six dimensions).

The result shows that most dimensions have not reached the Infusing stage, and a significant difference is presented among the levels of education, especially between secondary and primary school. The majority of dimensions have only reached the Emerging and Applying stage.

Therefore, with the goal of new technology-based schools, the transformative model of UNESCO (2005)

should be rearranged according to the school's relevant criteria in Vietnam, divided into stages comprised of (1) emerging, (2) applying, (3) infusing, and (4), transforming, specified as follows:

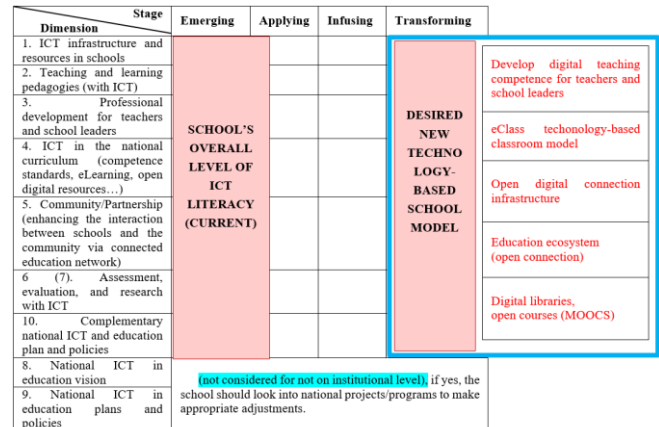


Fig. 8 An educational transformative sustainability model

This model focuses on 8 dimensions which are direct related to schools, in which the school's current overall level of ICT literacy is considered the first stage, and the new technology-based school model is the final goal. The pathway from emerging to transforming stage is presented below.

### C. Transformative pathway to technology-based school in modern educational technology-based approach

Given the characteristics of educational technology, the transformative pathway is developed based on the school's ICT literacy in reality, comprised of the following stages:

- **Stage 1 and 2:** focusing on developing teachers' and students' ICT competence on the grounds that ICT is used efficiently with teaching and learning pedagogies. This is a preparation step to form a habit and adaptability for teachers and students in the online environment.
- **Stage 3:** focusing on organizing technological blended classroom and implementing on the available system and infrastructure.
- **Stage 4:** evaluating and developing a suitable ecosystem for the school based on the outcome of stage 1, 2, and 3, moving toward to an open educational environment to ensure the sustainable operation. The details of each stage are presented in Table 3.

## III. CASE STUDY: HCMUTE DEVELOPS UTEX USING NEW TECHNOLOGY-BASED SCHOOL MODEL

### A. HCMUTE: Preparation for development

HCMUTE is one of the leading technology and education universities in Vietnam. With over 20,000 students in 14 Faculties/Institute and "humanity - innovation - integrity" as its educational philosophy, HCMUTE aims to innovate higher education on grounds of humanity, building an internationalized educational environment and encouraging creativity in education and entrepreneurship, consistent with the era 4.0 of digital knowledge.

TABLE 3. TRANSFORMATIVE PATHWAY TO TECHNOLOGY-BASED SCHOOLS

Stage Dimensions	Emerging Changing from the traditional	Applying Accepting the blended, diverse technology	Infusing Orienting integrity and synchronization	Transfor-ming Developing towards education ecosystem
Teaching and learning model with technology	Teaching and learning with effective, transformative ICT literacy	Teaching and learning with online, collaborative ICT literacy	Teaching and learning with technological blended classroom model	Teaching and learning with eClass model
School's ICT tools literacy	Using common ICT tools in classroom effectively	Using online, collaborative ICT tools extensively and optimally	Using ICT in technological blended classrooms flexibly and optimally	Using ICT in eClass flexibly and optimally
Pedagogical technology literacy	Teaching and learning effectively and actively with ICT tools	Changing to online teaching and learning optimally	Planning and organizing classroom in blended technology model	Planning and organizing eClass
Educational relationships (teachers, students, school leaders, parents)	Closed educational relationships	School-affiliated educational relationships (closed)	Open educational relationships (available)	<b>Open relationship education ecosystem</b>
ICT infrastructure	Network computing devices	Internet connection	Cloud (available)	Cloud-integrated school system
Information and learning system	Individual use	Open internet networks (available)	Blended learning and school management system	Personal system, digital library, eClass in the ecosystem
Establishing policies on technology-based school development	Short-term and small-scaled development policies	Supporting policies, accompanying the transforming and expanding stage	Supporting policies, accompanying the eClass approaching stage	Complete policies for long-term operation

A survey on the current status of digital education at HCMUTE was conducted. The results show that the school has made elaborate preparations on many aspects, among

which creating a digital education ecosystem is one of the focused long-term strategies to be developed, specified in Table 4 as follows:

TABLE 4. OVERVIEW OF DIGITAL EDUCATION AT HCMUTE

Aspects	2014 - 2018	2019	2020
1. The school's viewpoint on ICT	From the orientation to developing ICT and eLearning	Orientation to completing and boosting eLearning toward the Ecosystem	

## 2. Lecturers using ICT in teaching

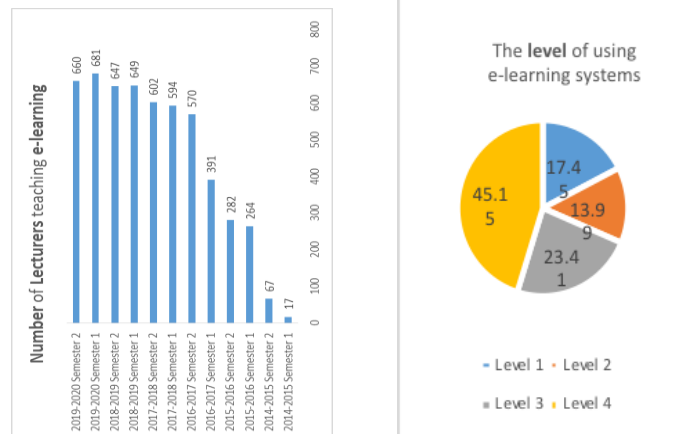


Fig. 9 Lecturers and online learning

### Levels of application

- + Level 1: learning materials
- + Level 2: materials + video
- + Level 3: materials + video + test
- + Level 4: materials + video + test + interaction with students on forums

Up to 82.55% of lecturers provide sufficient learning materials and lectures, and conduct tests on the school's LMS/FHQLMS

### 3. Number of online learning students participation & satisfaction

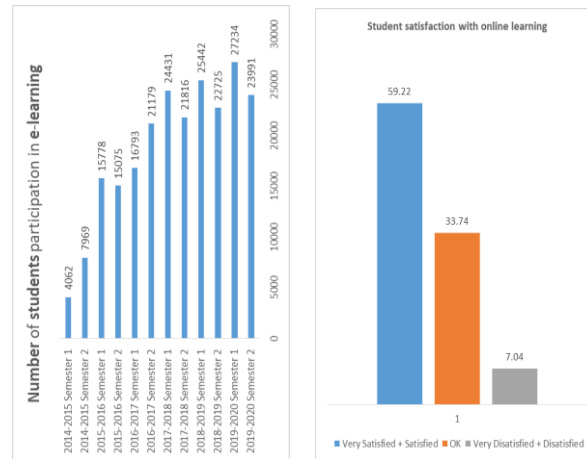


Fig. 10 Students and online learning

On average, there was 59.22% of students feel satisfied and very satisfied, 33.74% normal, and 7.04% dissatisfied and very dissatisfied.

4. Educational systems and infrastructure	2013: Pearson system provided by the University of Arizona for general programs (MoodleLMS) 2018: Blackboard sponsored high-quality programs (FHQLMS) 2019 Developed MOOCs (Edx) & established virtual learning center UTEX → oriented towards a future education ecosystem		
	School server	Upgraded school's server and data center	Oriented towards Cloud (Server)
5. Professional development training (research and training for digital learning and educational technology)	Digital learning training for teachers Available digital learning center	E-learning training for teachers Available Studio and educational technology center	Developing towards education ecosystem
6. Implementing the policy: <b>Carrot and stick</b>	SPECIFICATION: + Cash incentive (Level 2: 10 million, level 3: 20 million) + Research credit conversion (80, 120) + Certificates + Number of periods reduction + Training for digital learning		Supplementation and adjustments for the next phase

#### B. Evaluation of status quo and adjustment of solution for developing new technology-based school at HCMUTE

Based on the survey data, the proposed pathway is applied to the case study of HCMUTE.

The transformation into new technology-based school at HCMUTE shows that it requires a lengthy pathway with a lot

of determination and preparations to achieve the results so far, in the orientation to a future education ecosystem with a sustainable basis provided by modern educational technology. In the following overview (Table 5), the shaded areas represent the results that HCMUTE has or is about to achieve, while the white areas represent future goals.

TABLE 5. STATUS QUO FOR THE TRANSFORMATION INTO NEW TECHNOLOGY-BASED SCHOOL AT HCMUTE

Stage Dimensions	Emerging Changing from the traditional	Applying Accepting the blended, diverse technology	Infusing Orienting integrity and synchronization	Transforming Developing towards education ecosystem
Teaching and learning model with technology	From 2014 to 2018		From 2019, 2020 onwards	
	Achieving positive results when switching to online teaching and learning		<b>Orienting towards eClass model in the education ecosystem</b>	
School's ICT tools literacy	ICT competence is diverse; forming a collaborative culture in digital learning is necessary		Adapting flexibly and using ICT optimally in eClass	
Pedagogical technology literacy	Only exploiting ICT tools and not utilizing them as their essence of an instructional design		<b>Planning and organizing eClass in the education ecosystem</b>	
Educational relationships (teachers, students, school leaders, parents)	Having wide and international educational connections		Unlimited digital connections and open education	
ICT infrastructure	Expanding the infrastructure and switching to reliable Clouds		<b>Transforming the Cloud-based system with the school's available infrastructure</b>	
Information and learning system	Having its own information and learning system (derived from LMS 2.0)		Developing towards open education ecosystem	
Development policy	Carrot and stick policy		<b>In need of a policy for open education</b>	

It is noticeable that the key issues in the next phase are transforming in the orientation to educational technology and creating an education ecosystem. Therefore, the following imperatives should soon be dealt with:

1. Building a complete organizational structure with modern educational technology literacy to develop dimensions of new technology-based school
2. Building MOOCs on the new technology-based school model to develop and normalize teachers and students... to gradually shape the culture in this environment | **towards the digital literacy standards of ISTE** (International Society for Technology in Education) on campus-scale.
3. Building an education ecosystem corresponding to the digital education model, connected to open and

abundant resources... towards personalized lifelong learning.

4. Making timely adjustments to policies and regimes... facilitating the development of open educational ecosystem
5. Building a complete corresponding management and education quality assessment system in the education ecosystem.

With timely supplementation and adjustments of the aforementioned next steps, the successful transformation into a new education model, towards creating an open education ecosystem for sustainable development at HCMUTE is indisputable.

#### IV. CONCLUSION

The status quo at HCMUTE demonstrates the pathway to an educational transformative sustainability model based on modern educational technology, acting as a clear and detailed guideline which schools can easily refer to and use to evaluate their own status quo, thereby adjusting the solution to develop a new technology-based school consistent with a specific context. By understanding the essence of modern educational technology, the application of such into adjusting the transformative model consistent with the reality in Vietnam has increased education efficiency in the new context. This is also the impetus for many schools to extend the research towards unique education ecosystems, the combination of which would create a characteristic education ecosystem of Vietnam.

#### REFERENCES

- [1] Rieckmann, M. (2017). Education for sustainable development goals: Learning objectives. UNESCO Publishing.
- [2] Peña-López, I. (2015). Rethinking Education. Towards a global common good?
- [3] Leicht, A., Heiss, J., & Byun, W. J. (2018). Issues and trends in Education for Sustainable Development (Vol. 5). UNESCO Publishing.
- [4] Buckler, C., & Creech, H. (2014). Shaping the future we want: UN Decade of Education for Sustainable Development; final report. UNESCO.
- [5] Tran, T., Pham, T. X., & Vu, T. T. T. (2019, October). E-Class Education Model in Modern Educational Technology-Based Approach. In International Conference on Information, Communication and Computing Technology (pp. 405-416). Springer, Cham.
- [6] Sims, R. (2008). Rethinking (e) learning: A manifesto for connected generations. Distance Education, 29(2), 153-164.
- [7] Hase, S., & Kenyon, C. (2000). From andragogy to heutagogy. Ulti-BASE In-Site.
- [8] Hase, S., & Kenyon, C. (2003). Heutagogy and developing capable people and capable workplaces: strategies for dealing with complexity.
- [9] Hase, S., & Kenyon, C. (2007). Heutagogy: A child of complexity theory. Complicity: An international journal of complexity and education, 4(1).
- [10] Malek, J. A. (2018). The Impact of Heutagogy Education through Telecentre in Smart Village (SV). E-Bangi, 14(2).
- [11] Anderson, J. (2010). ICT transforming education: A regional guide. Published by UNESCO Bangkok, 120.
- [12] Engida, T. (2011). ICT-enhanced teacher development model. UNESCO-IICBA.
- [13] Jude, L. T., Kajura, M. A., & Birevu, M. P. (2014). Adoption of the SAMR model to asses ICT pedagogical adoption: A case of Makerere University. International Journal of e-Education, e-Business, e-Management and e-Learning, 4(2), 106.
- [14] Conesa, J., Batalla-Busquets, J. M., Bañeres, D., Carrion, C., Conejero-Arto, I., Gil, M. D. C. C. ... & Monjo, T. (2019, November). Towards an Educational Model for Lifelong Learning. In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 537-546). Springer, Cham.
- [15] Ilomäki, L., & Lakkala, M. (2018). Digital technology and practices for school improvement: innovative digital school model. Research and practice in technology enhanced learning, 13(1), 25.
- [16] United Nations Educational, Scientific and Cultural Organization (UNESCO). (2011). UNESCO ICT competency framework for teachers.
- [17] Southeast Asian Ministers of Education Organization (2010). Report of the Status of ICT integration in Education in South East Asian countries. [https://www.seameo.org/SEAMEOWeb2/images/stories/Publications/Project\\_Reports/SEAMEO\\_ICT-Integration-Education2010.pdf](https://www.seameo.org/SEAMEOWeb2/images/stories/Publications/Project_Reports/SEAMEO_ICT-Integration-Education2010.pdf) last accessed April 4, 2020.
- [18] UNESCO (2005). Regional Guidelines on Teacher Development for Pedagogy-Technology Integration. Thailand, Bangkok: UNESCO Asia and Pacific Regional Bureau for Education.
- [19] Kumar, K. L. (1996). Educational technology. New Age International.
- [20] Sebastian, K.: Educational Technology and Curriculum. 1st edn. ED-Tech Press, United Kingdom (2019).
- [21] VVOB Vietnam. Report on survey ICT in education status and target for 2020 (2012) [https://vietnam.vvob.org/sites/vietnam/files/report\\_on\\_survey\\_ict\\_in\\_education\\_status\\_and\\_targets\\_for\\_2020\\_v0.0\\_120418\\_vn.pdf](https://vietnam.vvob.org/sites/vietnam/files/report_on_survey_ict_in_education_status_and_targets_for_2020_v0.0_120418_vn.pdf) last accessed April, 8, 2020.



# A Novel Binning Algorithm Using Topic Modelling and k-mer Frequency on Groups of Non-Overlapping Short Reads

Hoang D. Quach

*Faculty of Information Technology  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
hoangqd@hcmute.edu.vn*

Dang H. N. Nguyen

*Faculty of Computer Science and Engineering  
HCMC University of Technology,  
Vietnam National University Ho Chi Minh City  
dangnguyenngochai.nnhd@gmail.com*

Hoang T. Lam

*Faculty of Computer Science and Engineering  
HCMC University of Technology,  
Vietnam National University Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
lamthanhhoang97@gmail.com*

Phuong V. D. Van

*Faculty of Computer Science and Engineering  
HCMC University of Technology,  
Vietnam National University Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
phuong@lhu.edu.vn*

Van Hoai Tran

*Faculty of Computer Science and Engineering  
HCMC University of Technology,  
Vietnam National University Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
hoai@hcmut.edu.vn*

**Abstract**—Metagenomics is a field that studies the microorganisms from the environment itself instead of traditional culturing methods. In this paper, we focus on the binning problem, which is to group reads into clusters that highly represent a taxonomic group. The result of this step serves as a crucial input for the next one of a metagenomic project such as assembly and annotation. Because metagenomic reads do not have explicit features, it is not easy to divide them into distinct groups. The solutions for this binning problem can be categorized as supervised and unsupervised approaches. Supervised ones need a reference database, which is unfortunately about 1% of the microorganisms in nature. This prevents these approaches from working well with the dataset that contain unknown species. In this paper we follow an unsupervised approach. Our proposed method is to combine the result from another technique named BiMeta, which based on a biological signature assumption that reads of a same taxonomic label have a same k-mer distribution, and topic modelling as a way of reducing the dimensions of the dataset. Our method shows better results (by *precision*, *recall*, and *F-measure*) than BiMeta on most datasets. Although following BiMeta, LDABiMeta outperforms it with the new proposed ideas. Moreover, our method is equivalent to MetaProb, which is the most successful method at present time, for the short-read datasets.

**Index Terms**—Metagenomic binning, topic modelling, short reads.

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2019-20-06.

## I. INTRODUCTION

One of the key targets in microbiological community research is to understand their composition, diversity and function. In the past, these problems were primarily solved by applying laboratory culture and cloning to the sequence of a specific gene because of its reasonable price and fast post-treatment. However, with the advancement of new generation sequencing techniques, the main focus has shifted to studying the whole metagenome shotgun taken directly from environment. This allows for more detailed analysis of metagenomic data, including the reproduction of the genomes of a new bacteria or virus, and even being able to gain knowledge of the genetic and metabolic potentials of the entire environment. However, the output of the new sequencing technologies is a mixture of short DNA fragments belonging to different genomes which requires more complex computational algorithms for clustering of related DNA fragments. This challenge is also the target of the first crucial phase in a metagenomic project, named *metagenomic binning* or *taxonomic binning*, in which a large set of DNA fragments (also called *reads*) is needed to separate into taxonomy-related groups. The output of the binning phase is helpful in assembly and annotation.

In metagenomics, where input is a genetic sequence, the two most important aspects that affect accuracy are sequence length and reference data. Computing time is second challenge, especially when working with large metagenomic data

sets. With a fast evolution of viruses and bacterial in nature, a complete reference database is not available for taxonomy-dependent methods. Therefore, *unsupervised approach* turns to be a reasonable choice, which depends strongly on the extraction of internal structure and information of genomic fragments to support clustering. However, this problem cannot be solved efficiently due to the limited length of the DNA fragments in binning.

Topic modelling has emerged as an effective method to explore hidden structures in a dataset. This paper only focuses on the application of topic modelling to metagenomic data, especially for the binning problem. When applying a topic model to metagenomic data, most studies have represented string data (DNA sequences) as documents, consisting of a set of  $k$  mers (or  $k$ -grams). Then, topic models are applied to analyze these set of words. In other words, each DNA sequence can be featured by a frequency vector, but now in topic space, instead of  $l$ -mers space.

In the study, we will introduce two new contributions. The first is the introduction of the frequency vector in the hidden topic space (topic distribution vector) into the clustering process. Using LDA as the core technique, the proposed approach will introduce some advanced techniques for turning a read into a document. However, purely applying this method, it is not easy to achieve high cluster accuracy. The new method combines with the technique of connecting reads by an overlapping graph, which bases on the similar idea in [1], [2]. With this support, clustering based on the topic distribution obtained from topic modelling will reduce noise, leading to increased accuracy even with the short read datasets.

The paper is organized as follows. Section II introduces related research, mainly focusing on the group of taxonomy-independent classification and also the latest research on the application of topic modelling in metagenomics. The next section presents a new proposal, which presents a 2-step approach which integrates LDA analysis. The computational results will be introduced in Section IV, in which we experiment on the same data sets used in other relevant papers. Then, the last section gives some concluding remarks on the applicability of the proposed approach.

## II. RELATED WORKS

Mande et al. [3] have a comprehensive review of binning methods for metagenomic data but mainly focused on methods based on a reference database. The readers who are interested in taxonomy-dependent methods refer to the latest review by Breitwieser et al. [4]. However, recently the taxonomy-independent binning is of great interest to the research groups to propose various methods. In this section, we describe the relevant studies in the second approach. Although these strategies do not require reference databases, they often have a higher requirement for sequence (read) length.

One common problem which is shared by all taxonomy-independent methods is *feature extraction*. There are two main types of features in binning process: feature based on *sequence composition* and feature based on *abundance* (or *coverage*).

Based on those two features, taxonomy-independent methods are separated into 3 types: (i) *sequence composition-based methods*, (ii) *abundance-based methods* and (iii) *hybrid methods*.

Composition-based methods assume that the composition of a genome is unique for each taxon. Therefore, it is possible to cluster reads by comparing their internal composition. Because a nucleotide sequence represents a read, the widely-used step to extract its composition information is to transform it into a numeric vector, called *genomic signatures*. A common way of vector transformation is  $l$ -mers frequency distribution, with a given  $k$ . For example, with  $k = 4$ , we have a vector of length 256 ( $4^4 = 256$ ). This vector shows the frequency of “words” ( $l$ -mers) with the length of 4 {AAAA, AAAC, ..., TTTT}. Based on the basic idea above, many studies have analyzed in-depth and proposed many advanced methods. The length of the distribution vector is its dimension limiting the application of composition-based methods to large data sets. In order to solve this challenge, some studies are focusing on reducing the dimension. Some of well-known ones are TETRA [5], Likely-Bin [6], SCIMM [7], VizBin [8], BiMeta [1], MetaProb [2].

A problem of composition-based methods is to cluster species with a low abundance. In more detail, reads belonging to small groups are wrongly assigned into groups of other species with higher abundance. This problem can be solved by using abundance-based methods. There are two main approaches to those methods. Some methods (Abundance-Bin [9], MBBC [10]) work with one metagenomic sample, and the other methods work with several metagenomic samples (Canopy [11]). The main idea of the abundance-based methodology is that the distribution of sequences follows the Poisson distribution, so the sequences are modeled as a mixture of Poisson distributions [9]. In other words, cluster formation is determined by the abundance of  $l$ -mers ( $l \geq 20$ ), instead of their similarities as in composition-based methods. Another problem with composition-based methods is that they only provide reasonably accurate results only when longer sequences are used (for example,  $\geq 800$ bp). Meanwhile, AbundanceBin [9] can work correctly even with sequences that are only 75bp long.

Hybrid methods use both sequence composition and abundance, often resulting in higher accuracy. CompostBin [12] is a pioneering method in this direction. Firstly, it extracts the frequencies of the 6-mers and uses principal component analysis (PCA) to reduce the number of dimensions. The method assigns weight to those frequencies based on the inverse value of abundance. Another method of using PCA to reduce dimensionality is CONCOCT [13]. CONCOCT combines two vectors:  $l$ -mers frequency and coverage within the PCA analysis. This method is only useful when the number of metagenomic samples is larger than 50bp. CONCOCT uses the Bayesian variational approach with the Gaussian mixture model (GMM) to estimate cluster numbers. Similar to CONCOCT, COCACOLA [14] combines coverage across multiple samples with genomic signatures to create feature vectors and uses  $L_1$  norm to calculate distances instead of Euclidean

measurements. Methods in this direction are MaxBin [15], MetaBAT [16], MyCC [17].

As mentioned in Section I, *natural language processing* (NLP) is gaining traction today. The interesting point is that four nucleotides (A, C, G, T) can be considered as an alphabet, which is an important basic component in NLP. Only those that use the NLP theory and techniques in metagenomics are reviewed. An overview on the application of topic modelling to bioinformatics is represented in [18]. For metagenomic binning, Chen and his colleagues used frequency vectors of  $l$ -mers to represent DNA sequences and applied the LDA model to infer hidden topics in the expectation that each hidden topic represents a particular gene [19]–[21]. La Rosa et al. took a similar approach, but in this study, hidden topics discovered in metagenomic data not only have a probability distribution on words ( $l$ -mers) but also correspond to a correct classification label (taxonomic label) (see [22]). With the LDA-based approach, Zhang et al. [23] used the SKWIC algorithm, a variant of the  $k$ -means algorithm, to group DNA sequences represented by hidden topics. According to these studies, the use of a topic model for metagenomic classification yields promising results when compared to other methods such as AbundanceBin [9], MetaCluster [24], and MCluster [25].

However, the challenge of the short length of the sequence affecting the binning's accuracy has not been solved in the end. The next section of the paper will present a method to address this challenge.

### III. LDABiMETA: A NOVEL COMBINATION OF LDA AND BiMETA

As mentioned above, a strategy that can provide good clustering quality must ensure that information within the DNA sequences is connected to the feature vector. The approach proposed in this paper includes two main ideas. Instead of working on the  $l$ -mers distribution vector, we recommend using the LDA (Latent Dirichlet Allocation) topic analysis tool to create the distribution vector on the topic. It can be said that the new approach is similar to the one proposed in MetaProb [2]. In this paper, the authors proposed a fairly complex probabilistic sequence signature, in which the  $l$ -mers distributions were carefully standardized at the global and local levels. This can be done automatically by topic modelling, and that is why we recommend using LDA. With the ability of a statistical generative process, LDA's analyzes will not only reduce the number of dimensions of the feature vector but also explore a topic space, which will help bring out a better feature vector in a hidden space, including global and local considerations. Note that with NLP, hidden topics often contain a label or a meaning, while with metagenomics, these topics do not have a meaning. So in the paper, we still use the phrase "hidden topic". Figure 1.a visualizes reads of dataset R4 (also used in [1], [2]), and colors corresponds to true read label. It is easy to see in the picture that reads is quite separate into two groups (red and black), and if there is an appropriate clustering method, the high accuracy of binning is feasible.

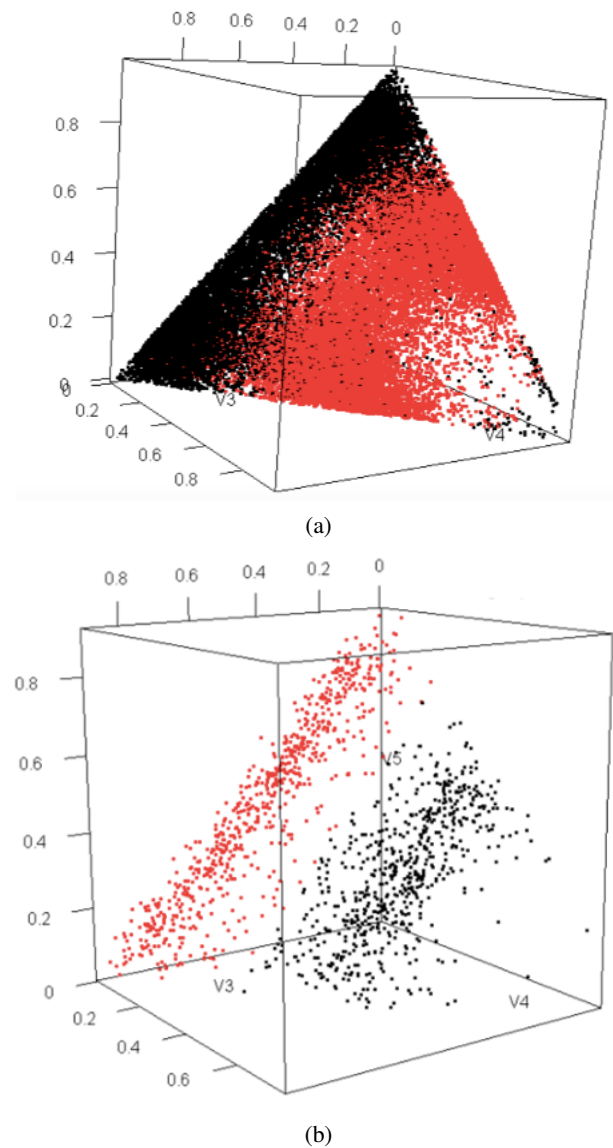


Fig. 1. 3D T-SNE visualization of dataset R4, LDA analysis with #topics = 3. (a) All reads (b) Only representative read groups

However, as seen in Figure 1.a, the contiguity between the two regions will cause many difficulties for clustering algorithms. In other words, the choice of parameters of clustering algorithms will be a significant obstacle, especially to be suitable for many different data sets. Therefore, we propose to use the second idea not perform clustering on every read. Instead, the new proposal will perform a partitioning of the read set first, based on the idea of the overlapping graph, creating groups. After that, only non-overlapping reads in the groups are used to aggregate the feature vector to represent the group. This approach can significantly reduce noise, caused by short reads that are quite similar to those of short reads but belong to other genomes. The second idea is illustrated in Figure 1.b, in which the red and black nodes have been separated, making clustering algorithms more efficient.

Based on the two ideas explained above, we propose a new method called LDABiMeta, shown in Figure 2. Firstly, we

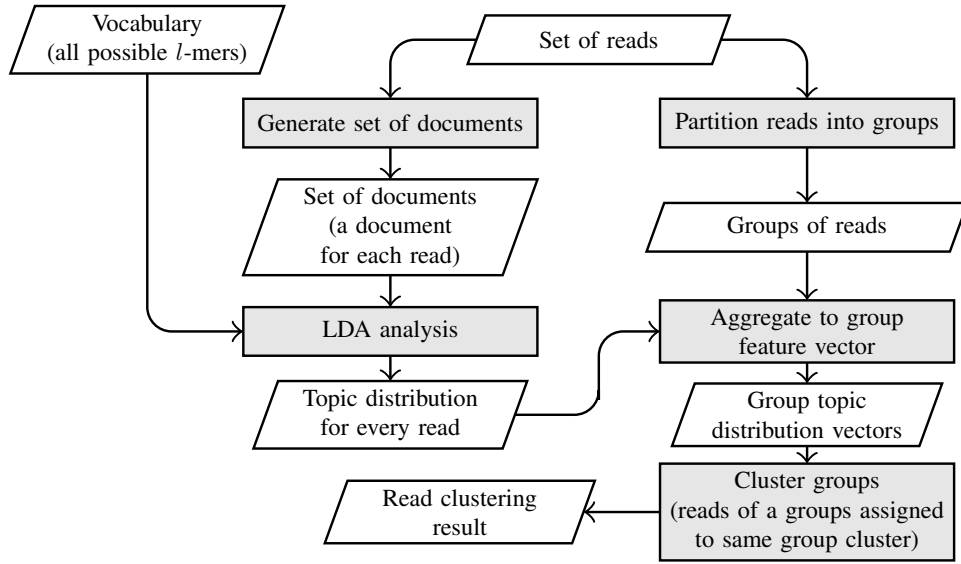


Fig. 2. LDABiMeta: a combination of LDA and BiMeta for short-read datasets.

represent each read as a document of successive  $l$ -mers. Then, we apply the LDA model on these documents to determine the hidden topic distribution for each read. The result of this phase is that each read is represented by a distribution vector of topics and each topic is represented by a distribution vector of  $l$ -mers. The weight of these vectors represent the contribution of each topic/ $l$ -mers to each read/topic, respectively. These distribution vectors will not be taken to the clustering phase immediately. Instead, another parallel process will create a set of groups along with a set of representative reads. Then, the topic distributions of the representative reads are aggregated into a topic distribution to representing the group. Clustering will be performed on these vectors. Note that when we partition reads into groups, it is assumed that that reads belonging to the same group would be in the same cluster. Therefore, after the clustering results for the group is available, we need to assign the cluster value to each read in the group to obtain the final clustering results.

The following sections will detail main algorithmic blocks in the proposed solution.

#### A. LDA analysis for metagenomic data

The main goal of LDA is to discover the hidden structure within a corpus. Figure 3 presents the main structure of LDA, using plate notation.

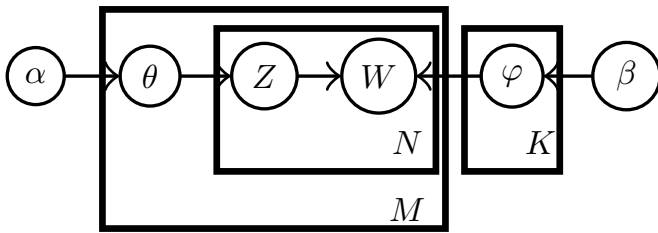


Fig. 3. The LDA Model [26]

In Figure 3,  $\alpha$  and  $\beta$  are two hyperparameters with Dirichlet prior, corresponding to the per-document topic distribution and the per-topic word distribution. With LDA analysis, only variable  $W$  is observable, relating to the set of words, while all others are hidden. Understandably, we can see  $\theta$  and  $\phi$  as variables that will be generated by LDA analysis (reverse of generative process), providing corresponding the per-document topic distribution vector and per-topic word distribution vector. To apply in this study, we must have a set of all possible words (also called *dictionary*) and a set of documents (also called *corpus*).

A metagenomic dataset consists of reads from different species. In order to apply the LDA model to metagenomic data, we have to find a way to represent the reads (strings) similar to the documents (set of words). In NLP, a dictionary is needed in LDA analysis. However, it is not available in metagenomics, so this study has proposed a dictionary creation by creating all  $l$ -mers from 4 nucleotides (A, C, G, T) in which  $l$  is given. The same method also applies to document creation from reads, where each  $l$ -mers is a substring of  $l$  consecutive characters in the read. For example, if we have read  $r = \text{ATCGAAGGTCGT}$  and choose  $l = 4$  then the  $l$ -mers substrings of  $r$  are ATCG, TCGA, CGAA, GAAG, AAGG, AGGT, GGTC, GTCG, and TCGT. Thus, read  $r = \text{ATCGAAGGTCGT}$  will be expressed in document as “ATCG TCGA CGAA GAAG AAGG AGGT GGTC GTCG TCGT”.

Once the document set is available, any topic model can be used for analysis to detect the underlying structure. More specifically, instead of using the  $l$ -mers distribution as in other studies [1], [2], the LDA model was used to create topic distributions for each read, which could be applicable in the clustering phase.

### B. Finding non-overlapping reads

If the clustering is carried out immediately based on the topic distribution generated by the LDA analysis presented in the previous section, the result will probably be unexpectedly low because of the noise caused by short reads. Instead, we apply the idea in [1] to separate the reads into groups. Our approach is quite similar to [1], [2], but the main difference is that we have generalized this stage into two sub-problems: graph partitioning and representative set finding. From there, we can apply different partitioning methods and find representative sets.

In this study, we still reuse the graph creation method proposed in [1]. The reads will be compared against each other based on similar measurements, also reusing the idea of  $n$ -grams. Specifically, an  $l$ -mers distribution will be created for each read, and as verified in [1], if  $l$  is large enough ( $l \geq 20$ ), this  $l$ -mers distribution is enough to represent a genome. With the overlapping graph, our study still uses the heuristic approach based on the two related studies to make fair assessments. However, after we have read groups, we do not use the simple heuristic way to find the representative set as in [1], [2]. Instead, we apply a method of finding approximately the stable sets, which could give larger size representative sets.

### C. The combination of LDA and BiMeta

We combine the read groups from BiMeta and topic distributions of every read from LDA to create feature vectors for each group. Let  $\{G_i\}_i$  are the read groups, created from BiMeta,  $t_j, j \in \{1, 2, \dots, n\}$  is the topic distribution of read  $r_j$ , created from LDA. The pseudocode for LDABiMeta, a reality of this combination, is showed at Algorithm 1.

---

**Algorithm 1: LDABiMeta**


---

**Input:**  $G = \{G_i\}_i, R = \{r_j\}_j, T = \{t_j\}_j$

**Output:** Labeled reads  $\{r_j\}_j$

- 1 Create representative vectors for each group  $G_i$  as  
 $v_{G_i} = \text{mean}(t_j), \forall r_j \in G_i$
  - 2 Clustering  $v_{G_i}$  using  $k$ -means to create cluster  
 $C_h, h \in \{1, \dots, p\}$
  - 3 **for**  $h \leftarrow 1$  **to**  $p$  **do**
  - 4     1. Find all the groups  $G$  such that  $v_G \in C_h$
  - 5     2. Assign all the reads in group  $G$  with the label  $h$
  - 6 **end for**
- 

## IV. EXPERIMENTS

### A. Data

The data used for performance evaluation are the ones used by the study of Vinh et al. [1]. These are datasets generated by MetaSim [27]. In this study, we focus on experimenting with short sequences because they are the most to cluster precisely. The short sequence datasets are about 80bp in length. These datasets have error rates of about 1% and are described in detail in Table I.

TABLE I  
SHORT-READS METAGENOMIC DATASETS (ADAPTED FROM [1] AND [2])

Dataset	No. of species	Phylogenetic distance	No. of reads (ratio)
S1	2	Species	96367 (1:1)
S2	2	Species	195339 (1:1)
S3	2	Order	338725 (1:1)
S4	2	Phylum	375302 (1:1)
S5	3	Species and Family	325400 (1:1:1)
S6	3	Phylum and Kingdom	713388 (3:2:1)
S7	5	Order, Order, Genus, Order	1653550 (1:1:1:4:4)
S8	5	Genus, Order, Order, Order	456224 (3:5:7:9:11)
S9	15	various distances	2234168 (1:1:1:1:1:1:2:2:2:2:2:3:3:3:3:3)
S10_S	30	various distances	1500000 (4:4:4:4:4:6:6:6:6:6:7:7:7:7:7:8:8:8:8:8:9:9:9:9:9:10:10:10:10:10)
L1	2	Class	176688 (1:1)
L2	2	Class	259568 (1:2)
L3	2	Class	342448 (1:3)
L4	2	Class	425328 (1:4)
L5	2	Class	508209 (1:5)
L6	2	Class	591089 (1:6)

### B. Evaluation measures

Let a metagenomic dataset  $D$  consisting of  $n$  reads  $x_i$ , partitioned into  $p$  species. Let  $y_i \in \{1, 2, \dots, p\}$  denote the ground-truth labels (species) for each read. The ground-truth clustering is given as  $T = \{T_1, T_2, \dots, T_p\}$ , where the cluster  $T_j$  consists of all the reads with label  $j$ , i.e.,  $T_j = \{x_i \in D | y_i = j\}$ . Also, let  $C = \{C_1, C_2, \dots, C_k\}$  denote a clustering of the same dataset into  $k$  clusters, obtained via some clustering algorithm, and let  $\hat{y}_i \in \{1, 2, \dots, k\}$  denote the cluster label for  $x_i$ . We will refer to  $T$  as the ground-truth partitioning, and to each  $T_i$  as a partition. We will call  $C$  a clustering, with each  $C_i$  referred to as a cluster. Because the ground truth is assumed to be known, typically clustering methods will be run with the correct number of clusters, that is, with  $k = p$ . However, to keep generality, we allow  $k$  to be different from  $p$ .

The clustering validation measures try capture the extent to which reads from the same partition (species) appear in the same cluster, and the extent to which reads from different partitions are grouped in different clusters. These measures rely on the  $k \times p$  contingency table  $N$  (see Table II) that is induced by a clustering  $C$  and the ground-truth partitioning  $T$ , defined as follows

$$N(i, j) = n_{ij} = |C_i \cap T_j|$$

TABLE II  
CONTINGENCY TABLE OF CLUSTERING RESULTS

Clusters/Species	$T_1$	$T_2$	$\dots$	$T_p$
$C_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1p}$
$C_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$C_k$	$n_{k1}$	$n_{k2}$	$\dots$	$n_{kp}$



The measures used to evaluate clustering in this study results are *precision*, *recall*, and *F-measure*. These measures are calculated from Table II using the below formulas.

$$precision = \frac{\sum_{i=1}^k \max_{j \in \{1, \dots, p\}} \{n_{ij}\}}{\sum_{i=1}^k \sum_{j=1}^p n_{ij}} \quad (1)$$

$$recall = \frac{\sum_{j=1}^p \max_{i \in \{1, \dots, k\}} \{n_{ij}\}}{\sum_{i=1}^k \sum_{j=1}^p n_{ij}} \quad (2)$$

$$F\text{-measure} = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (3)$$

### C. Experimental results

The study uses the library BioString<sup>1</sup> to preprocess the metagenomic data and to transform all datasets into documents of *l*-mers. LDA module in the library text2vec<sup>2</sup> is to detect hidden topics and to reduce *l*-mers distribution vector to topic distribution vector. A variant of BiMeta [1] is applied to find out non-overlapping reads of the groups that are generated by a greedy graph partitioning phase.

In order to evaluate the impact of LDA in dimensional reduction, we apply LDA analysis for every read. In those experiments, the number of topics is set to 10 for all datasets. We called this method LDACluster. The number of topics is a hyper-parameter and can be tuned to achieve better results.

To evaluate the effectiveness of the feature vector generated by LDA, we compared its effectiveness with a method based on *l*-mers distribution vector only. We call BaseCluster a method to classify metagenomic data based on the application of *k*-means on the *l*-mers distribution vector of reads. We have the table III summarizing *F-measure* between BaseCluster and LDACluster. The results in this table show that LDACluster is more effective than BaseCluster. Although the difference is not large, this table shows that the topic vector generated by LDA is more efficient than the *l*-mers distribution vector.

TABLE III  
*F-measure* OF BASECLUSTER AND LDACLUSTER.

Dataset	BaseCluster	LDACluster
S1	<b>0.767</b>	0.672
S2	0.582	<b>0.659</b>
S3	0.909	<b>0.936</b>
S4	<b>0.964</b>	0.963
S5	0.531	<b>0.544</b>
S6	<b>0.950</b>	0.945
S7	<b>0.485</b>	0.457
S8	0.450	<b>0.464</b>
S9	<b>0.364</b>	0.363
S10_S	<b>0.253</b>	0.208
L1	0.627	<b>0.701</b>
L2	0.634	<b>0.790</b>
L3	0.673	<b>0.817</b>
L4	0.692	<b>0.854</b>
L5	0.706	<b>0.910</b>
L6	0.714	<b>0.892</b>

However, LDACluster is not impressive because it does not take advantage of the biological characteristics of metagenomic data but mechanically applies the LDA model to this data. To overcome this situation, we sought to take advantage of the biological characteristics of metagenomic data into the model. In this experiment, we combined the ideas of BiMeta and LDA. Phase 1 of BiMeta helps to group highly similar readings into groups (using *l*-mers), and this has helped to make use of the biological characteristics of metagenomic data. Details of the processing steps are described in Section III. The paper compares the performance of LDABiMeta with state-of-the-art metagenomic binning algorithms: AbundanceBin [9], BiMeta [1], and MetaProb [2] in Table IV. It can be seen in Table IV, LDABiMeta works well with all short-read datasets. Except the large datasets S9 and S10\_S, LDABiMeta is quite stable in performance. In average, LDABiMeta is equivalent to MetaProb which is one of most successful metagenomic binning methods.

TABLE IV  
*F-measure* OF ABUNDANCEBIN, METACLUSTER, BIMETA, METAPROB (ADAPTED FROM [1], [2]), AND LDABIMETA ON SHORT-READ DATASETS.

Dataset	AbundanceBin	MetaCluster	BiMeta	MetaProb	LDABiMeta
S1	0.683	0.672	0.978	<b>0.991</b>	<b>0.991</b>
S2	0.713	0.631	0.581	0.901	<b>0.911</b>
S3	0.824	0.415	0.978	0.928	<b>0.983</b>
S4	0.883	0.46	<b>0.994</b>	0.908	0.993
S5	0.552	0.643	0.69	0.832	<b>0.856</b>
S6	0.692	0.492	0.858	0.97	<b>0.993</b>
S7	0.606	0.652	<b>0.843</b>	0.782	0.811
S8	0.528	0.529	0.743	<b>0.769</b>	0.758
S9	Error	0.639	<b>0.791</b>	0.719	0.635
S10_S	0.137	0.052	0.429	<b>0.495</b>	0.305
L1	0.625	0.549	0.98	<b>0.984</b>	0.983
L2	0.793	0.675	0.98	<b>0.992</b>	0.976
L3	0.9	0.667	0.986	<b>0.993</b>	0.989
L4	0.959	0.703	<b>0.987</b>	0.986	0.98
L5	0.977	0.612	<b>0.991</b>	0.983	0.984
L6	0.984	0.649	<b>0.99</b>	0.984	0.981
Average	0.723	0.565	0.862	<b>0.888</b>	0.883

<sup>1</sup><https://bioconductor.org/packages/release/bioc/html/Biostrings.html>

<sup>2</sup><http://text2vec.org/>

## V. CONCLUSION

The short length of reads is one of the main factors that reduce the performance of metagenomic binning. Because the information contained in a single read is not sufficient to correctly determine the same-cluster relationship with other reads. This challenge has been solved by the proposed method, called LDABiMeta. A new combination of LDA and the method for finding non-overlapping reads has been experimented on short read datasets. Compared to BiMeta, the proposed method outperforms in *F-measure* in general and *precision/recall* in particular for most datasets. This result proves that using topic distribution as the feature vector is a correct approach, and is also consistent with the research in MetaProb (using a probabilistic sequence signature). Compared with MetaProb, LDABiMeta has equal clustering quality, thereby showing its potential in metagenomic binning. As indicated in Section III, LDABiMeta still has many points that can be further studied, especially exploiting the research in NLP in general or topic modelling in particular in metagenomics.

**Acknowledgment.** This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2019-20-06.

## REFERENCES

- [1] LV Vinh, TV Lang, LT Binh, and TV Hoai. A two -phase binning algorithm using l-mer frequency on groups of non overlapping reads. *Algorithms for Molecular Biology*, 10(2), 2015.
- [2] S Giroto, C Pizzi, and M Comin. Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics*, 32(17):567–575, 2016.
- [3] TS Ghosh SS Mande, MH Mohammed. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6):669–681, 2012.
- [4] FP Breitwieser and SL Salzberg J Lu. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, 20(4):1125–1136, 2019.
- [5] H Teeling, J Waldmann, T Lombardot, M Bauer, and FO Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5(163), 2004.
- [6] A Kislyuk, S Bhatnagar, J Dushoff, and JS Weitz. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, 10(1), 2009.
- [7] D Kelley and S Salzberg. Clustering metagenomic sequences with interpolated markov models. *BMC Bioinformatics*, 11(544), 2010.
- [8] CC Laczny, T Sternal, V Plugaru, P Gawron, A Atashpendar, and H Margossian. Vizbin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, 3(1), 2015.
- [9] Y-W Wu and Y Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tupless. *Journal of Computational Biology*, 18(3):523–34, 2011.
- [10] Y Wang, H Hu, and X Li. Mbbc: an efficient approach for metagenomic binning based on clustering. *BMC Bioinformatics*, 16(1), 2015.
- [11] HB Nielsen, M Almeida, AS Juncker, S Rasmussen, and J Li. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–8, 2014.
- [12] S Chatterji, I Yamazaki, Z Bai, and J Eisen. Compostbin: A dna composition-based algorithm for binning environmental shotgun reads. *Lecture Notes in Computer Science*, 4955:17–28, 2008.
- [13] J Alneberg, BS Bjarnason, I De Bruijn, M Schirmer, J Quick, and UZ Ijaz. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–6, 2014.
- [14] YY Lu, T Chen, JA Fuhrman, and F Sun. Cocacola: binning metagenomic contigs using sequence composition, read coverage, co-alignment and paired-end read linkage. *Bioinformatics*, 33(6):791–798, 2017.
- [15] Y-W Wu, Y-H Tang, SG Tringe, BA Simmons, and SW Singer. Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2(1), 2014.
- [16] DD Kang, J Froula, R Egan, and Z Wang. Metabat - an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3(e1165), 2015.
- [17] H-H Lin and Y-C Liao. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep*, 2016.
- [18] L Liu, L Tang, W Dong, S YaoEmail, and W ZhouEmail. An overview of topic modeling and its current applications in bioinformatics. *Springer-Plus*, 5(1), 2016.
- [19] X Chen, X Hu, X Shen, and G Rosen. Probabilistic topic modeling for genomic data interpretation. in *IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2011.
- [20] X Chen, T He, X Hu, Y Zhou, and Y An et al. Estimating functional groups in human gut microbiome with probabilistic topic models. *IEEE Transactions on NanoBioscience*, 11(3):203–215, 2012.
- [21] X Chen, X Hu, TY Lim, and X Shen. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):980–991, 2012.
- [22] M La Rosa, A Fiannaca, R Rizzo, and A Urso. Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC Bioinformatics*, 16(S2), 2015.
- [23] R Zhang, Z Cheng, J Guan, and S Zhou. Exploiting topic modeling to boost metagenomic reads binning. *BMC Bioinformatics*, 16(5):1–10, 2015.
- [24] Yi Wang, Henry C.M. Leung, S.M. Yiu, and Francis Y.L. Chin. Meta-cluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics*, 28(18):356–362, 2012.
- [25] R Liao, R Zhang, J Guan, and S Zhou. A new unsupervised binning approach for metagenomic sequences based on n-grams and automatic feature weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(1):42–54, 2014.
- [26] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- [27] DC Richter, F Ott, AF Auch, R Schmid, and DH Huson. Metasim - a sequencing simulator for genomics and metagenomics. *PLoS ON*, 3(10), 2008.

# Study on the Influence of Diaphragm Wall on the Behavior of Pile Raft Foundation

Van Qui Lai\*  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology, VNU-HCM,  
Ho Chi Minh City, Viet Nam  
lvqui@hcmut.edu.vn

Quoc Thien Huynh  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology, VNU-HCM,  
Ho Chi Minh City, Viet Nam  
hqthienxdbk@gmail.com

Nhat Hoang Vo  
Faculty of Civil Engineering,  
Ho Chi Minh City University of  
Technology, VNU-HCM,  
Ho Chi Minh City, Viet Nam  
nhathoangvo0809@gmail.com

Chung Nguyen Van  
Faculty of Civil Engineering, HCMC University of Technology and Education.  
Ho Chi Minh City, Viet Nam  
chungnv@hcmute.edu.vn

**Abstract**—The paper focus on analyzing the influence of diaphragm wall on the bending moment and settlement of pile raft foundation. A case study in Ho Chi Minh City using pile-raft foundation was used for studying process. The Finite element method (FEM) with the Hardening Soil model (HS) were employed to describe the behavior of soils. Behavior of pile-raft foundation was investigated by considering the available of diaphragm wall and the connection between diaphragm wall and raft. The results showed that the settlements of pile-raft foundation were decreased 12-20% and the moments of the raft were increased 17-20% when considering the influence of diaphragm wall. The connection between diaphragm wall and raft did not effect on the settlement and moment in middle area of pile-raft foundation.

**Keywords**—Deep excavation, pile-raft foundation, Ho Chi Minh City

## I. INTRODUCTION

Nowadays, the speed of urbanization is growing faster, so more and more high-rise buildings have built to serve the needs of houses, commercial centers, offices. Following that, the foundations become more complex to ensure adequate bearing capacity for the project. The pile raft foundation has become an effective solution for foundation engineering. It has used for many large projects in the world such as: Euro Tower (148.5m), Dresdener (166.7m), Commerz Bank (108.5m), Westend (208m), Deutsche Bank (92.5 m), Main Tower (205m), BFG Bank (186m), Messertum (256.5m), City Bank (55m). The advantage of the raft foundation is the ability to withstand large loads, cost savings more than the plan of pile foundation or conventional raft foundation. There have been many studies on pile foundation [1-5]. However, these studies have mainly focused on methods of internal force analysis, calculation, and design of raft foundations. They were neglected factors affecting internal force of rafts and piles such as the working of the basement system with the raft or the working of the diaphragm wall with the raft.

In this paper, the authors analyze the influence of diaphragm wall on behavior of pile raft foundation. For particular problem, the paper has studied on different of internal forces, settlement of raft between the case considering diaphragm wall surrounding pile raft foundation and the case of absence of the diaphragm wall. Additionally, the stiffness of the connection between diaphragm wall and the raft, which are difficult to determine, was also considered in the study. A

case of pile raft foundation placed on thick sand layer in Ho Chi Minh City, was used for analyzing and investigation. The 3D Finite element method (FEM), namely 3D Plaxis, was used in the computational performance.

## II. PROJECT DESCRIPTION

The representative project using the pile raft foundation is investigated. The project is located in district 1, Ho Chi Minh City that has four basements. The soil profile was shown schematically in Figure 1.

THICKNESS	LAYERS	N(SPT)	DESCRIPTION
1.0	A	0	LAYER A : BACK FILL
3.9	1	1	LAYER 1st : sandy clay medium stiff
31.1	2	17	LAYER 2nd : Clayed sand - medium dense
14.3	3	30	LAYER 3rd : Sandy clay - stiff
7.5	4	36	LAYER 4th : Sandy clay - medium stiff
12.4	5	41	LAYER 5th : Clay sand - dense
6.9	6	56	LAYER 6th : Sandy clay - very stiff
> 42.9	7	79	LAYER 7th : Sand - very dense

Fig. 1. Soil profile of studied project

The pile raft foundation was designed with piles having a diameter of 1500mm in the middle area and 1200mm of the edge area. The raft thickness was 3.0m, and the foundation depth was 18m compared to the ground surface. The raft foundations were placed on the sandy ground with medium SPT index 17. The thickness of the diaphragm wall was 800mm and 40m long from the ground surface. Figure 2(a) was showed the detailed plan and Figure 2(b) was showed the investigated cross section of the pile raft foundation. Diameter piles (D1500) had the bearing capacity was 18000kN and D1200 piles was 10000kN. The actual design did not considered the influence of the diaphragm wall to the pile foundation, as shown in Figure 2.

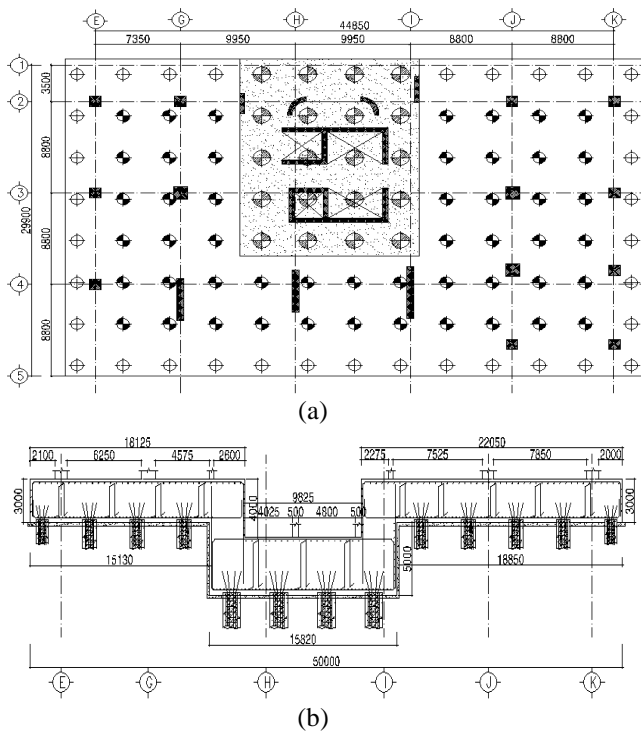


Fig. 2. a. Plan of the pile raft foundation; b. Cross section of the pile raft foundation

### III. STUDY ON THE INFLUENCE OF DIAPHRAGM WALL ON THE BEHAVIOR OF PILE RAFT FOUNDATION

#### A. Investigation cases

To investigate the influence of diaphragm on pile raft foundation, there was some cases proposed for analyzing. The first case, namely Case 00, did not consider diaphragm wall working with pile raft foundation. In this case, the diaphragm wall was absence in the Case 00. In the second case, namely Case 01, was considered diaphragm wall working with pile raft foundation. However, in this case was lacking the connection between the diaphragm wall and the raft. In fact, the stiffness of the connection between the raft and the diaphragm wall cannot be determined precisely. The steel linking between the diaphragm wall and the raft was implanted or already in-stalled. Then, the concrete connecting between the diaphragm wall and the raft did not pour in the same time with the diaphragm wall and the raft. Thus the connection between the diaphragm wall and the raft was a complex connection. In the other hand, determining the exact stiffness at this position was impossible. It was not fully investigated as it was neglected the connection between the diaphragm wall and the raft as Case 01 or it was considered the full stiffness of the connection. Therefore, in this study, influence of the connection on the behavior of pile raft foundation was studied by varying the stiffness of the connection. With 3D FEM model, the authors changed the stiffness of the connection by varying the cross-sectional area. In particularly, from Case 01 (no connection), to Case 07 (fully connection) the cross-sectional area of the connection was varied from 0 m thickness to 3m thickness, respectively. The all studied case was showed in Figure 3.

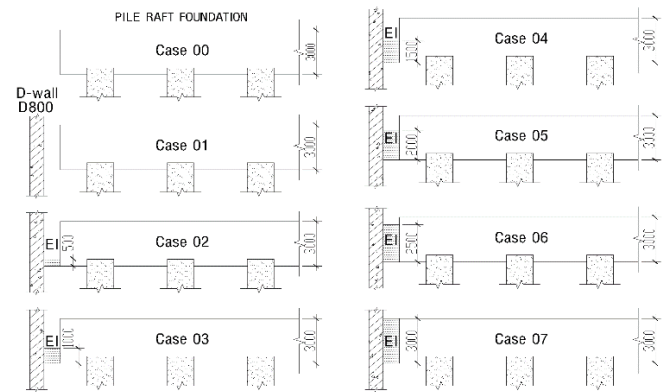


Fig. 3. Parameter study of stiffness of connection between pile raft foundation and D-wall

#### B. Soil, structure and 3D FEM model

The 3D FEM, namely Plaxis 3D, was used for the analysis purposes. Hardening soil model which was normally using in geotechnical analysis [6-8], Figure 4, was selected to analyze the behavior of soil layers. HS model parameters were determined from laboratory experiments such as CD, CU, OED, DS and field experiments according to semi-empirical methods such as VST, SPT.

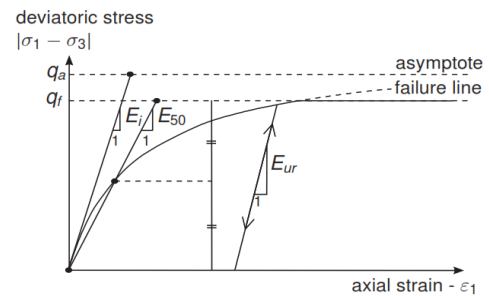


Fig. 4. Hardening soil model (Plaxis manual)

For cohesive soils, in the absence of a 3-axially CD or CU test, the total stress analysis (Undrained B in Plaxis) was used. The values of strength parameter included  $\phi_u = 0$ , the non-drainage shear strength  $c_u = S_u$  were determined from field or laboratory experiments. The value of  $E_{50}$  was taken from  $S_u$  value.

For cohesion-less soils, the effective stress analysis was used. Shear strength parameters were taken from direct shear (DS) test or correlated with SPT value ( $N_{SPT}$ ).  $N_{SPT}$  taken from field experiments was widely used in foundation design. For instance, Japanese Architectural Institute [9] proposed the formula for calculating the pile load capacity by  $N_{SPT}$ . For the stiffness value,  $E_{50}$ , can be taken according to  $N_{SPT}$ . In recent study,  $E_{50}$  was proposed to  $2000N_{SPT}$  [10]. Stroud [11] also presented the relationship between the values of  $E_{50}$  and  $N_{SPT}$  by collecting data from different soil types, and he suggested that  $E_{50}$  decrease as the strain decreases. Basing on the Japanese Architectural Institute recommended  $E = 2800N_{SPT}$  [9], the geological parameters in this study were shown in Table 1.

For structures of the pile raft foundation including piles, raft and diaphragm walls were modeled by element in Plaxis 3D. A plate element was used for flexural structures such as diaphragm walls, basements and rafts. The embedded pile element was used for pile. The input material parameters of structure elements were shown as in Table 2. The view of the

3D model was shown in Figure 5(a). Besides, the load at the foot of the column includes vertical force  $N$  and the moment  $M$  from upper structure model were directly imported according to the position of the column foot in Plaxis 3D model as Figure 5(b).

TABLE I. INPUT SOIL PARAMETER

Layers	Back fill	Layer 1: sandy clay, medium stiff	Layer 2: clayey sand, medium dense	Layer 3: sandy clay, stiff	Layer 4: sandy clay, medium stiff	Layer 5: clayey sand, dense	Layer 6: sandy clay, very stiff	Layer 7: Sand, very dense
Type	HS Drained	HS Un-drained	HS Drained	HS Un-drained	HS Un-drained	HS Drained	HS Un-drained	HS Drained
$\gamma_{sat}$ (kN/m <sup>3</sup> )	18	15.8	19.78	20.21	19.85	19.97	20.17	19.72
$\gamma_{sat}$ (kN/m <sup>3</sup> )	18.5	15.95	20.28	20.59	20.21	20.46	20.68	20.07
$E_{50}^{ref}$ (kN/m <sup>2</sup> )	15000	7500	33000	44000	56000	60000	70000	70000
$E_{50}^{ref}$ (kN/m <sup>2</sup> )	15000	7500	33000	44000	56000	60000	70000	70000
$E_{ur}^{ref}$ (kN/m <sup>2</sup> )	45000	22500	99000	132000	168000	180000	210000	210000
$m$	1	0.8	0.5	0.9	0.8	0.5	1	0.9
$\nu_{ur}$	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
$c'$ (kN/m <sup>2</sup> )	5	15	8.3	30.9	23.8	10.5	30.5	5.3
$\phi$ (°)	25°00'	25°51'	30°50'	25°47'	23°15'	31°25'	28°15'	31°20'
$\psi$ (°)	0	0	0°50'	0	0	1°25'	0	1°20'
$R_{inter}$	0.6	0.7	0.8	0.9	0.9	0.9	0.9	0.9
Level of layer bottom	-1.00	-5.00	-36.00	-50.00	-57.80	-70.20	-78.00	>100

TABLE II. PARAMETER OF FOUNDATION STRUCTURE.

Structures	Diaphragm wall	Basements	Raft	Piles
Thickness $t$ (m)	0.8	0.3	3	-
Diameter $D$ (m)	-	-	-	1.2-1.5
Unit weight $\gamma$ (kN/m <sup>3</sup> )	25.0	25.0	25.0	25.0
Module $E$ (Mpa)	27.0	32.5	32.5	27.0
Poisson's ratio	0.2	0.2	0.2	0.2

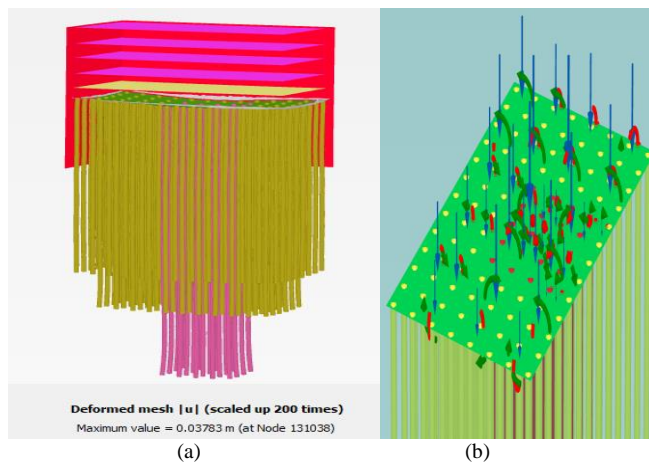


Fig. 5. 3D FEM model, (a). Pile raft foundation and D-wall, (b). Input external load

### C. Results

The firstly investigation was analyzed the settlement and internal moment of the pile raft foundation. As shown in

Figure 6 and Figure 7, the largest settlement at middle of the raft in Case 00 and Case 01 were 42.4mm and 37.26mm, respectively. This result indicated that the influence of the diaphragm wall reduced the settlement of the raft by 12%. Furthermore, it was shown that the influence of the diaphragm walls on the working of the raft foundation, the settlement of raft became more evenly, the largest deformation area at the center became smaller. It can be explained that the soil below raft compressed by the raft cannot move in horizontal way by the present of diaphragm walls. So that, the ground below the raft was increased the capacity and stiffer.

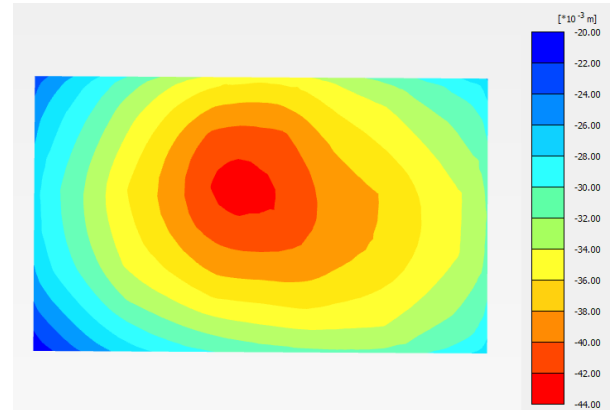


Fig. 6. Settlement of the pile raft foundation in Case 00

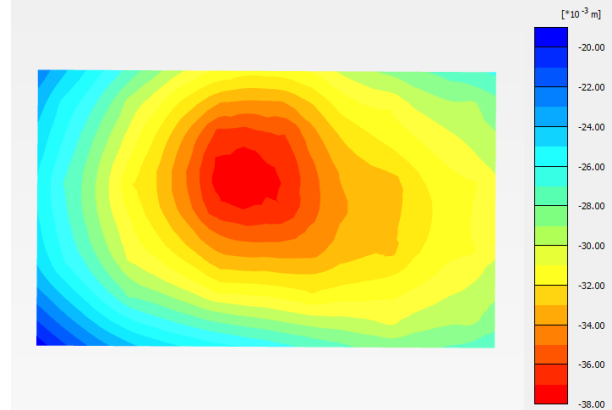


Fig. 7. Settlement of the pile raft foundation in Case 01

To investigate the influence of the stiffness of connection between the raft and diaphragm wall, the comparison between the settlements at the cross section 1-1 in all cases were shown in Figure 8 and Figure 9. When the thickness of the bonding increased 0 m to 3.0m, respect to Case 01 to Case 07, the settlement at the cross session did not change too much. Furthermore, the largest settlements at the middle area of Case 07 were smaller, around 2%, compare to Case 01. It proved that the connection between raft and diaphragm wall did not significantly affect on the settlement at the middle of the raft.

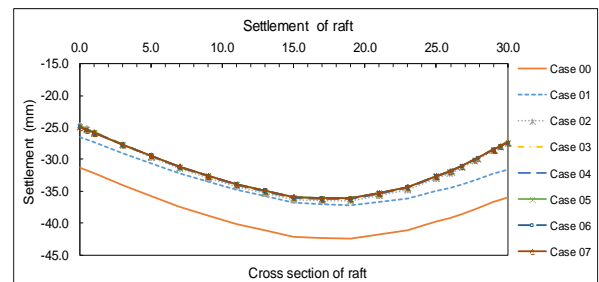


Fig. 8. Settlement of the pile raft foundation in all cases



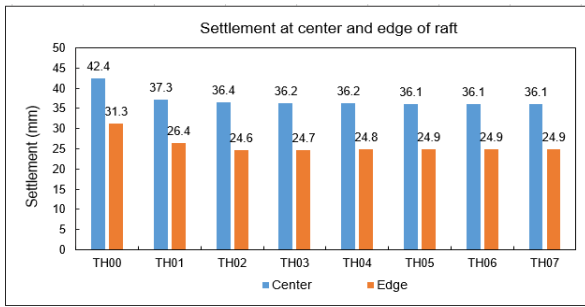


Fig. 9. Results of settlement at center and edge of raft in all investigated cases

Besides, considering the settlement at the edge of the raft, the largest settlement of Case 07 was 24.9mm and Case 00 is 31.3mm. The results showed that the settlement at the edge of the raft reduced 20% when considering influence of diaphragm wall. Therefore, it can be said that the settlement of pile raft was reduced 12%-20% by investigation throughout the cross section 1-1.

For internal moment investigation, the bending moment  $M_{22}$  at cross section 1-1 of Case 01 and Case 07 were shown in Figure 10. It can be seen that the moments in middle and edge area of the raft was increased. When the connections between raft and diaphragm wall was considered, the raft became stiffer. Thus, the axially load from upper structure was distributed on the raft more than on the pile. So that, in Case 07, the raft would be more bending than Case 01 having free connection between the raft and the diaphragm wall. Besides, the moment also appeared in fixing connection between raft and diaphragm wall in Case 07.

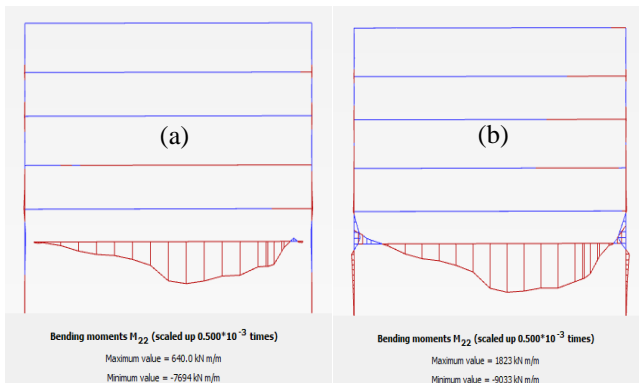


Fig. 10. Internal moment of raft : a. Case 01, b. Case 07

For more investigation, the moment at cross section 1-1 for all cases were shown in Figure 11 and Figure 12. In Case 00, absence of the diaphragm wall, and Case 01, absence of the connection between the diaphragm wall and the raft, the moment at cross section 1-1 of the raft was similar. It meant that the diaphragm wall having free connection with the raft did not effect on the moment of the raft. However, there was a huge difference in considering the connection between the diaphragm wall and the raft. Specifically, the moment at middle and edge in Case 07 increased 17% and 20%, respectively, compared to the moment at middle in Case 01 or Case 00. Besides, the stiffness of the connection was only effect on moment at edge. Similar to the previous investigation, the moment at middle from Case 02 to Case 07 were the same and moment at edge increased when stiffness of the connection between the diaphragm wall and the raft was increased.

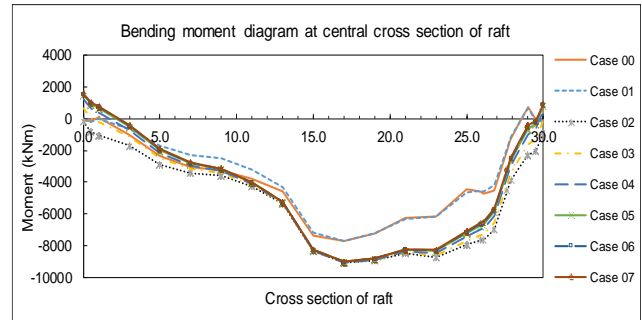


Fig. 11. Internal moment of the raft in all cases

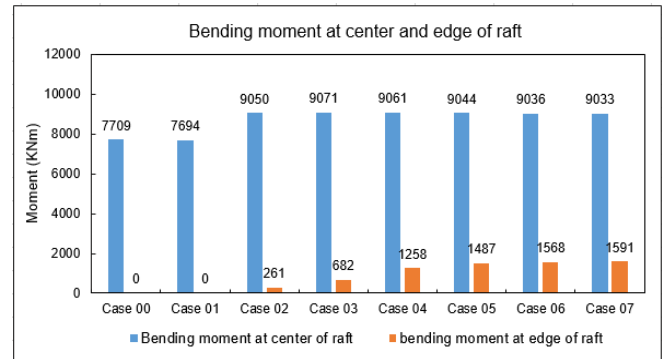


Fig. 12. Internal moment at center and the edge of raft in all investigated cases

For more significant investigation, the comparisons of the moments and the largest settlements at middle area of all cases were explored as shown in Figure 13. It can be seen that the largest settlements of raft were decreased from Case 00 to Case 07. Inversely, the moments at middle area were increased from Case 00 to Case 07. However, the moments from Case 02 to Case 07 were the same increment compared to Case 00.

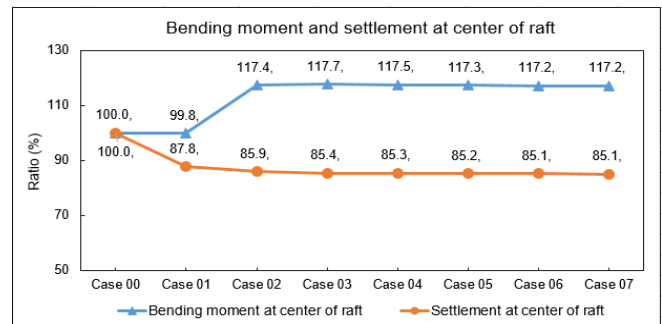


Fig. 13. Comparison between settlement and internal moment in the center of the raft in all cases

#### IV. CONCLUSION

The paper presented an investigation of the influence of diaphragm wall on the behavior of pile raft foundation by 3D FEM. A case study in the thick layers of sand in Ho Chi Minh City was selected to model and analyze. Several conclusions can be drawn:

- The settlement of the raft foundation was reduced by 12% - 20% when the influence of the diaphragm wall was considered. The stiffness of the connection between raft and diaphragm wall did not significantly affect to the settlement of raft.

- In the case of free connection between the raft and diaphragm wall, for Case 01, the diaphragm wall did not affect the moment of the raft. In the cases of considering connection between the raft and diaphragm wall, for Case 07, the moment in middle and edge area were increased 17% and 20%, respectively, compared to moment in middle area of Case 00 or Case 01.
- Connection between the raft and diaphragm wall only affected on moment values at boundary position.
- The stiffness of the connection between the diaphragm wall and the raft did not affect on moment and settlement at middle position of raft. The stiffness of the connection only affected on the moment at the edge of the raft.

#### REFERENCES

- [1] R. Katzenbach, U. Arslan, CHR. Moormann. Design and Safety concept for piled raft foundationn. Proc. 3rd Intl. Geotech seminar on deep foundations on bored and Auger piles, 19-25 October, 1998.
- [2] HG. Poulos HG. Piled raft foundations design and applications. *Géotechnique*, 51(2), 95-113, 2001.
- [3] DDC. Nguyen, SB. Jo, DS. Kim. Design method of piled-raft foundations under vertical load considering interaction effects. *Computers and Geotechnics*, 47, 16-27, 2013.
- [4] L. Zhang L, SH. Goh, J. Yi. A centrifuge study of the seismic response of pile-raft systems embedded in soft clay. *Géotechnique*, 67(6), 479-490, 2017.
- [5] Y. Liu, L. Zhang. Seismic response of pile-raft system embedded in spatially random clay. *Géotechnique*, 69(7), 638-645, 2019.
- [6] QT. Huynh, VQ. Lai, VT. Tran, MT. Nguyen. Analyzing the settlement of adjacent buildings with shallow foundation based on the horizontal displacement of retaining wall. *Lecture Notes in Civil Engineering*, 62, 313-320, 2020.
- [7] VQ. Lai, MN. Le, QT. Huynh, TH. Do. Performance analysis of a combination between D-wall and Secant pile wall in upgrading the depth of basement by Plaxis 2D: A case study in Ho Chi Minh city, *Lecture Notes in Civil Engineering*, 80, 2020.
- [8] QT. Huynh QT, VQ Lai, VT Tran, MT. Nguyen. Back analysis on deep excavation in the thick sand layer by hardening soil small model, *Lecture Notes in Civil Engineering*, 80, 2020.
- [9] AIJ Japan, Recommendations for design of building foundations, 2001.
- [10] BCB. Hsiung, KH. Yang, W. Aila, L. Ge. Evaluation of the wall deflections of a deep excavation in Central Jakarta using three-dimensional modeling. *Tunnelling and Underground Space Technology*, 72, 84-96, 2018.
- [11] M. Stroud. Penetration Testing in the UK, Thomas Telford, London, 1989

# A Theoretical and Numerical Study of Ultrasonic Waves in Laminated Composites for Nondestructive Evaluation of Structures

Duy Kien Dao

Faculty of Civil Engineering  
Ho Chi Minh city of Technology and  
Education  
Ho Chi Minh City, Vietnam  
kiendd@hcmute.edu.vn

Ductho Le

Faculty of Mathematics, Mechanics and  
Informatics  
VNU University of Science  
Hanoi, Vietnam  
ducthole24@gmail.com

TruongGiang Nguyen

Institute of Mechanics  
VAST  
Hanoi, Vietnam  
ngtrgiang10@gmail.com

Hoai Nguyen

Institute of Physics  
VAST  
Hanoi, Vietnam  
hoai.ngo@gmail.com

Duc Chinh Pham

Institute of Mechanics  
VAST  
Hanoi, Vietnam  
pdchinh@imech.vast.vn

Haidang Phan <sup>1,2</sup>

<sup>1</sup>Graduate University of Science and  
Technology, VAST  
<sup>2</sup>Institute of Theoretical and Applied  
Research, Duy Tan University  
Hanoi, Vietnam  
phanhaidang2@duytan.edu.vn

**Abstract**— The current work investigates guided waves in laminated composite plates subjected to ultrasonic sources for potential applications in nondestructive evaluation of composite structures. The propagation of free ultrasonic waves in laminated plates is first considered and their dispersion curves are obtained using numerical computation. The problem of wave motion generated by a time-harmonic load in the composites is theoretically derived in a simple manner using reciprocity theorems. Analytical and numerical results of the generated wave fields are compared, and they show excellent agreement for several material combinations. Scattering of guided waves by a delamination at the interface of laminates is also discussed. The expressions and calculations found in this investigation are critical to explore the potential of using ultrasound-based methods for nondestructive evaluation of composite structures.

**Keywords** — Guided waves, Nondestructive evaluation, Composite structures, Reciprocity, Delamination

## I. INTRODUCTION

Composite materials are being widely used in various components of engineering structures such as automotive parts, civil infrastructures, and especially aerospace components. The aerospace usage of composites has experienced a continuously growing over several decades. As an example, the Airbus A350 XWB contains up to 83% of total composite materials by volume and 52% by weight [1]. Aerospace composite structures are made by overlapping several unidirectional fiber-reinforced composite layers with different angle orientations. These composite materials have orthotropic material properties as they are much stiffer along the fiber direction than across the transverse direction.

Guided waves can propagate in composite plates with little loss of energy and inspect large areas from a single location thus they are critical to the applications of nondestructive evaluation (NDE) and structural health monitoring (SHM) of aerospace structures. Understanding of wave interaction with composites is still, however, far from complete due to the nature of multi-layer and anisotropic behaviors [2]. The main

purpose of this paper is to use the reciprocity theorem for direct calculation of the motion of guided waves in laminate composites due to a time-harmonic load. The dispersion equation is presented resulting in dispersion curves. The closed-form solutions of scattered amplitudes of guided waves in laminated composite plates are derived through the reciprocity relations between an actual state, wave motions generated by a time-harmonic line load, and a virtual state, a free guided wave chosen appropriately. These calculations can be considered as novel results since they have not been reported in literature. The analytical predictions presented in here helps provide better understanding of guided waves in composite structures and improve the methods of solving inverse scattering problems for applications in nondestructive evaluation.

The study of free guided wave propagation in unidirectional composites can be found, for example, in textbooks [2-4] and in numerous research articles, see [5, 6] for example. The wave motion due to ultrasonic sources in the composite laminate plates is much more complicated but unquestionably one of the most fundamental problems of elastodynamics. Conventionally, wave motion generated by a loading is solved using integral transform technique [7]. However, it is more difficult to use this classical approach for anisotropic materials and impossible to apply the method to the wave motion in inhomogeneous media. In order to avoid these complications, another analytical approach based on reciprocity in elastodynamic, strictly to determine the guided waves, has been proposed in recent years, see [8-12]. The applications of elastodynamic reciprocity theorems in order to obtain favorite information about the scattered field of guided waves generated by ultrasonic loads in layered structures were studied in [13-17]. Verification of the reciprocity approach used in the current study was presented in [11, 18, 19] for the case of surface waves.

In elastodynamics, reciprocity theorems offer an integral relation between two loading states of an elastic body [8]. The relation can be used to derive systems of integral equations. In

the current work, we aim to suggest direct applications of the reciprocity theorems to compute guided wave fields in a unidirectional fiber-reinforced composite plate subjected to a time-harmonic load for further development of ultrasound-based NDE and SHM techniques. By appropriately choosing the virtual state as a free guided wave, the scattered amplitudes of guided waves generated by the load (the actual state) are obtained through the reciprocity relations

## II. GUIDED WAVES IN A UNIDIRECTIONAL LAMINA

A thin unidirectional lamina of thickness  $2h$  in the Cartesian coordinate system  $(x, y, z)$  is illustrated in Fig. 1. The material is orthotropic. Consider the propagation of guided waves in the lamina along the  $x$ -direction with the phase velocity  $c$  and the wave number  $k$ .

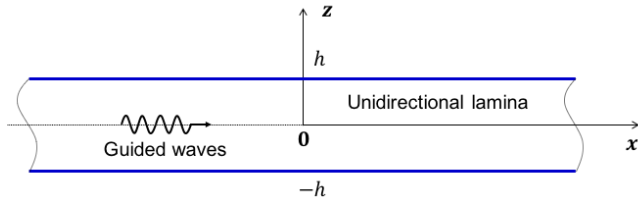


Fig. 1. Dispersion curves of unidirectional fiber composite with  $0^\circ$  fiber, Lamb type guided waves (solid), Shear horizontal waves (dash)

The displacement components in the  $(x, z)$  plane may be written in the forms of

$$u_x = (A_1 e^{is_1 kz} + A_2 e^{is_2 kz} + A_3 e^{-is_1 kz} + A_4 e^{-is_2 kz}) e^{ik(x-ct)} \quad (1)$$

$$u_z = (\alpha_1 A_1 e^{is_1 kz} + \alpha_2 A_2 e^{is_2 kz} - \alpha_1 A_3 e^{-is_1 kz} - \alpha_2 A_4 e^{-is_2 kz}) e^{ik(x-ct)} \quad (2)$$

where  $A_1, A_2, A_3, A_4$  are unknown constants and  $t$  indicates time. In Eqs. (1) and (2), we define

$$\alpha_j = -\frac{(C_{12} + C_{66})s_j}{C_{66} - \rho c^2 + C_{22}s_j^2} = -\frac{C_{11} - \rho c^2 + C_{66}s_j^2}{(C_{12} + C_{66})s_j}, \quad j = 1, 2 \quad (3)$$

where

$$s_1 = \sqrt{\frac{S + \sqrt{S^2 - 4P}}{2}}, \quad s_2 = \sqrt{\frac{S - \sqrt{S^2 - 4P}}{2}} \quad (4)$$

with

$$S = -\frac{(C_{11} - \rho c^2)C_{22} + (C_{66} - \rho c^2)C_{66} - (C_{12} + C_{66})^2}{C_{22}C_{66}} \quad (5)$$

$$P = \frac{(C_{11} - \rho c^2)(C_{66} - \rho c^2)}{C_{22}C_{66}} \quad (6)$$

In Eq. (3),  $\rho$  is the mass density and  $C_{11}, C_{12}, C_{22}, C_{66}$  are four elastic constants. It is noted that for axisymmetric or

asymmetric deformations, in general, there are five constants. For plane-strain deformations, only four constants are involved since the governing equations are independent of the material constant  $C_{13}$ . Based on the stress-strain relations for an orthotropic medium, stress components are computed as

$$\sigma_{xx} = ikc_{66}[\delta_1 A_1 e^{is_1 kz} + \delta_2 A_2 e^{is_2 kz} + \delta_1 A_3 e^{-is_1 kz} + \delta_2 A_4 e^{-is_2 kz}] e^{ik(x-ct)} \quad (7)$$

$$\sigma_{xz} = ikc_{66}[\beta_1 A_1 e^{is_1 kz} + \beta_2 A_2 e^{is_2 kz} - \beta_1 A_3 e^{-is_1 kz} - \beta_2 A_4 e^{-is_2 kz}] e^{ik(x-ct)} \quad (8)$$

$$\sigma_{zz} = ikc_{66}[\gamma_1 A_1 e^{is_1 kz} + \gamma_2 A_2 e^{is_2 kz} + \gamma_1 A_3 e^{-is_1 kz} + \gamma_2 A_4 e^{-is_2 kz}] e^{ik(x-ct)} \quad (9)$$

where

$$\delta_j = \frac{C_{11}}{C_{66}} + \frac{C_{12}}{C_{66}} \alpha_j s_j, \quad \beta_j = \alpha_j + s_j, \quad (10)$$

$$\gamma_j = \frac{C_{12}}{C_{66}} + \frac{C_{22}}{C_{66}} \alpha_j s_j, \quad j = 1, 2$$

## III. GUIDED WAVES IN LAMINATED COMPOSITE PLATE

A sketch of a laminated composite plate which includes several orthotropic layers is shown in the Fig. 2. Since  $h^{(n)}$  is the thickness of each layer, the thickness of the plate is  $H = \sum_{n=1}^N h^{(n)}$ . Besides the global coordinate  $(x, z)$ , it is convenient to also use a local coordinate system  $(x^{(n)}, z^{(n)})$  for layer  $n$ .

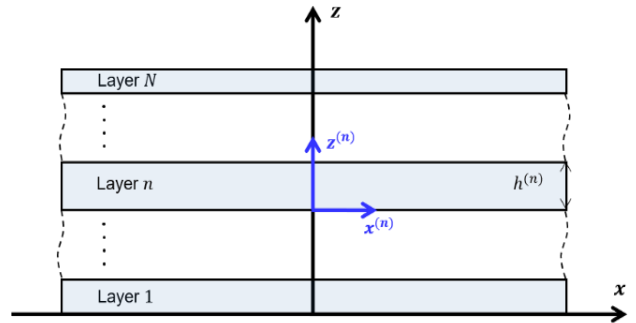


Fig. 2.  $N$ -layered laminated composite plate

We can write wave field solutions of displacement and stress components of the layer  $n$  ( $n = 1, 2, \dots, N$ ) as

$$\mathbf{Y}^{(n)} = [u_x^{(n)} \quad u_z^{(n)} \quad \sigma_{xz}^{(n)} \quad \sigma_{zz}^{(n)}]^T = \mathbf{\Gamma}^{(n)} e^{ik(x-ct)} \quad (11)$$

where

$$\mathbf{\Gamma}^{(n)} = \mathbf{X}^{(n)} \mathbf{W}^{(n)} \mathbf{A}^{(n)} \quad (12)$$

with

$$\mathbf{X}^{(n)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ \alpha_1^{(n)} & \alpha_2^{(n)} & -\alpha_1^{(n)} & -\alpha_2^{(n)} \\ ikc_{66}^{(n)}\beta_1^{(n)} & ikc_{66}^{(n)}\beta_2^{(n)} & -ikc_{66}^{(n)}\beta_1^{(n)} & -ikc_{66}^{(n)}\beta_2^{(n)} \\ ikc_{66}^{(n)}\gamma_1^{(n)} & ikc_{66}^{(n)}\gamma_2^{(n)} & ikc_{66}^{(n)}\gamma_1^{(n)} & ikc_{66}^{(n)}\gamma_2^{(n)} \end{bmatrix} \quad (13)$$

and

$$\mathbf{W}^{(n)} = \begin{bmatrix} e^{iks_1^{(n)}z^{(n)}} & 0 & 0 & 0 \\ 0 & e^{iks_2^{(n)}z^{(n)}} & 0 & 0 \\ 0 & 0 & e^{-iks_1^{(n)}z^{(n)}} & 0 \\ 0 & 0 & 0 & e^{-iks_2^{(n)}z^{(n)}} \end{bmatrix} \quad (14)$$

and

$$\mathbf{A}^{(n)} = [A_1^{(n)} \quad A_2^{(n)} \quad A_3^{(n)} \quad A_4^{(n)}]^T \quad (15)$$

The characteristic equations may be found by either global matrix method (GMM), transfer matrix method (TMM) or stiffness matrix method (SMM). Curious readers can refer to the textbooks by Giurgiutiu [1], Rose [2] and Nayfeh [20] for more details. For this study, a computer program is made based on the transfer matrix method to obtain dispersion curves of phase and group velocities.

At the bottom of the  $n^{th}$  layer where  $z^{(n)} = 0$ , we have

$$\mathbf{W}_0^{(n)} = \mathbf{I} \quad (16)$$

where  $\mathbf{I}$  is the identity matrix of order four and  $\mathbf{W}_0^{(n)}$  indicate the value of  $\mathbf{W}^{(n)}$  at the bottom of layer  $n$  where  $z^{(n)} = 0$ . Thus,

$$\mathbf{\Gamma}_0^{(n)} = \mathbf{X}^{(n)} \mathbf{A}^{(n)} \quad (17)$$

At the top of the  $n^{th}$  layer

$$\mathbf{\Gamma}_h^{(n)} = \mathbf{X}^{(n)} \mathbf{W}_h^{(n)} \mathbf{A}^{(n)} \quad (18)$$

$$\mathbf{\Gamma}_h^{(n)} = \mathbf{X}^{(n)} \mathbf{W}_h^{(n)} (\mathbf{X}^{(n)})^{-1} \mathbf{\Gamma}_0^{(n)} = \mathbf{\Psi}^{(n)} \mathbf{\Gamma}_0^{(n)} \quad (19)$$

with

$$\mathbf{\Psi}^{(n)} = \mathbf{X}^{(n)} \mathbf{W}_h^{(n)} (\mathbf{X}^{(n)})^{-1} \quad (20)$$

Boundary conditions

$$\mathbf{\Gamma}_h^{(n)} = \mathbf{\Gamma}_0^{(n+1)}, \quad n = 1, 2, \dots, N-1 \quad (21)$$

Then we have

$$\mathbf{\Gamma}_0^{(n+1)} = \mathbf{\Psi}^{(n)} \mathbf{\Gamma}_0^{(n)} \quad (22)$$

We obtain

$$\mathbf{\Gamma}_h^{(N)} = \mathbf{\Psi} \mathbf{\Gamma}_0^{(1)} \quad (23)$$

with the overall transfer matrix is defined by

$$\mathbf{\Psi} = \prod_{n=1}^N \mathbf{\Psi}^{(n)} \quad (24)$$

We have

$$\mathbf{Y}_h^{(N)} = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} & \Psi_{14} \\ \Psi_{21} & \Psi_{22} & \Psi_{23} & \Psi_{24} \\ \Psi_{31} & \Psi_{32} & \Psi_{33} & \Psi_{34} \\ \Psi_{41} & \Psi_{42} & \Psi_{43} & \Psi_{44} \end{bmatrix} \mathbf{Y}_0^{(1)} \quad (25)$$

Considering the stress-free condition at the bottom of the first layer and the top of the  $N^{th}$  layer results in two separate equations

$$\begin{bmatrix} u_x^{(N)} \\ u_z^{(N)} \end{bmatrix}_h = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix} \begin{bmatrix} u_x^{(1)} \\ u_z^{(1)} \end{bmatrix}_0 \quad (26)$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \Psi_{31} & \Psi_{32} \\ \Psi_{41} & \Psi_{42} \end{bmatrix} \begin{bmatrix} u_x^{(1)} \\ u_z^{(1)} \end{bmatrix}_0 \quad (27)$$

Nontrivial solution

$$\begin{vmatrix} \Psi_{31} & \Psi_{32} \\ \Psi_{41} & \Psi_{42} \end{vmatrix} = 0 \quad (28)$$

For nontrivial solutions, the determinant is required to be zero that results in a complicated equation with an unknown phase velocity. This equation can only be solved by numerical methods which results in dispersion curves. For guided waves propagating in a laminated composite plate which includes several unidirectional layers with different angle orientations, note that we can rotate their stiffness matrices and consider the wave problems in the same coordinate system.

Since Lamb type wave modes and shear horizontal (SH) wave modes are inseparable in general anisotropic materials, our computer code is intended for solutions of all guided wave modes. Therefore, SH wave modes will be appeared in the representation of results although they are not of interest in the current work. For the case study, we use material properties of IM7/997-3 CFRP which has a density  $\rho = 1560 \text{ kg/m}^3$  and a stiffness matrix given in Eq. (29). Examples of phase and group velocity solutions are shown in Fig. 3 (unidirectional fiber composite with  $0^\circ$  fiber) and Fig. 4 (four-layered laminated composite  $[0/45/90/135]$ ).



$$\mathbf{C} = \begin{bmatrix} 178.2 & 8.347 & 8.347 & & & \\ 8.347 & 14.44 & 8.119 & & & \\ 8.347 & 8.119 & 14.44 & & & \\ & & & 3.161 & & \\ & & & & 6.1 & \\ & & & & & 6.1 \end{bmatrix} \quad (29)$$

Here we can see for the case of 0 degree lamina, the Lamb type waves are separated to the SH waves. However, when the fiber angles are different with 0 and 90, Lamb type waves and SH waves are not decoupled.

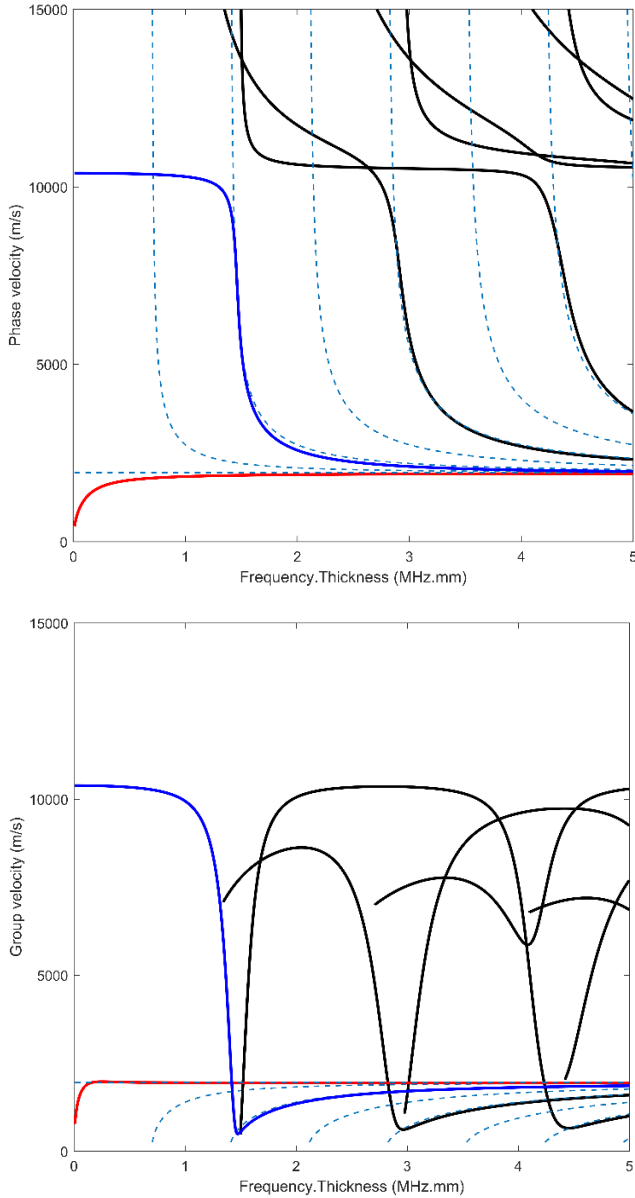


Fig. 3. Dispersion curves of unidirectional fiber composite with 0° fiber, Lamb type guided waves (solid), Shear horizontal waves (dash)

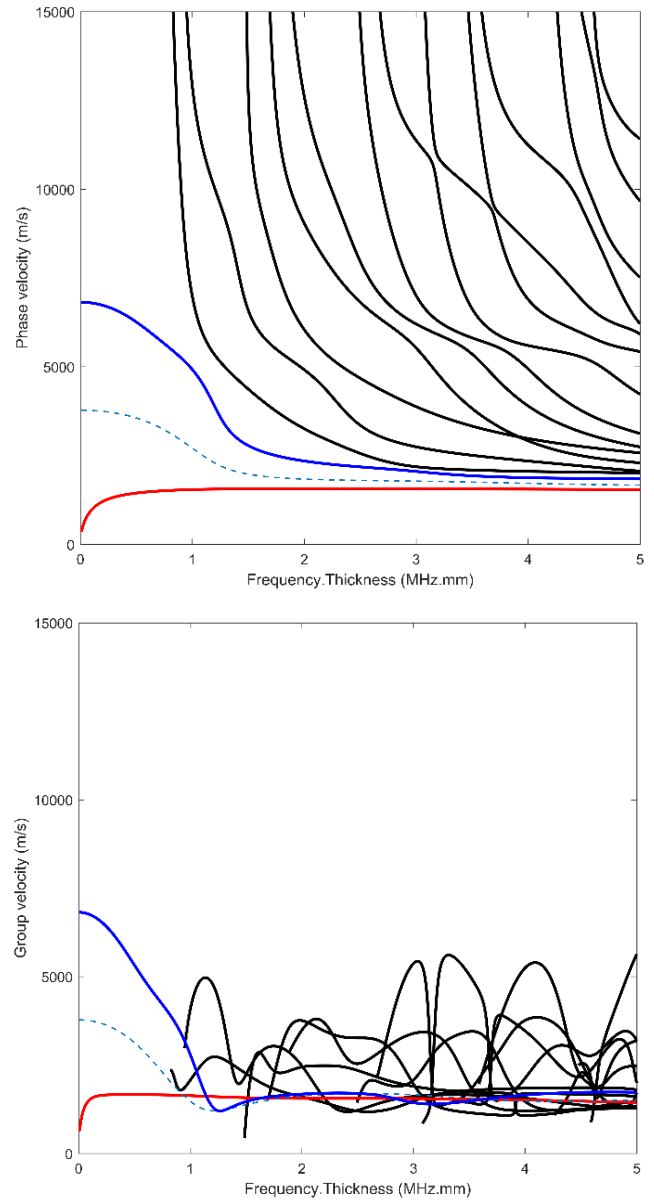


Fig. 4. Dispersion curves of multi-layer laminated composite [0/45/90/135], Lamb type guided waves (solid), Shear horizontal waves (dash)

#### IV. GUIDED WAVE MOTIONS IN A COMPOSITE PLATE BY RECIPROCITY CONSIDERATIONS

Reciprocity theorem is a relation between displacement, traction components and forces of two different loading states of a body that may be expressed in the form of [8]

$$\int_V (f_j^A u_j^B - f_j^B u_j^A) dV = \int_S (\tau_{ij}^B u_j^A - \tau_{ij}^A u_j^B) n_i dS \quad (30)$$

where  $S$  defines a contour around an area defined by  $V$ . Here, the superscripts  $A$  and  $B$  indicate two elastodynamic states,  $u_j$  and  $\tau_{ij}$  denote the displacements and stresses, respectively, while  $f_j$  is the force term and  $n_i$  represents the outward normal vector.

In case of a multi-material domain including multiple bonded bodies, a reciprocity theorem can be derived as

$$\sum_{l=1}^L \int_{V_l} (f_j^A u_j^B - f_j^B u_j^A) dV_l \quad (31)$$

$$= \int_{S_l} (\tau_{ij}^B u_j^A - \tau_{ij}^A u_j^B) n_i dS_l$$

where  $S_l$  ( $l = 1, 2, \dots, L$ ) are closed curves around  $V_l$  without the interface with other bodies.

In the present work we concern ourselves with applications of reciprocity to elastodynamic states to compute scattered wave field due to time-harmonic sources. The approach is based on the explicit expressions of free guided wave fields in composites. They can be in either analytical or numerical forms. In Eq. (31), state  $A$  is called the actual state, i.e., actual waves generated by a load while state  $B$  is called the virtual state, a free guided wave traveling in composites. By choosing a suitable virtual state and substituting it into Eq. (31), the actual wave solution due to the loading can be found. The reciprocity approach is shown to be much simpler than the conventional integral transform method and it can be used for anisotropic and inhomogeneous materials.

Let us consider an  $N$ -layered plate in a global coordinate system  $(x, z)$  given in Fig. 2. It is convenient to define the local coordinate system  $(x^{(n)}, z^{(n)})$  for layer  $n$ .

For free guided waves propagating in multilayered plates, global matrix method (GMM), transfer matrix method (TMM) and stiffness matrix method (SMM) are commonly used to obtain the characteristic equations [1]. In this study based on the calculation by GMM, numerical forms of the displacement components are found.

Computation of guided wave fields subjected to a time-harmonic vertical load given in Eq. (32) is now considered.

$$f_z^A = P \delta(z - z_0) \delta(x - x_0) e^{-ikct} \quad (32)$$

The load will generate antisymmetric guided wave modes with scattered amplitudes  $A_{p+}^A$  and  $A_{p-}^A$  in the  $x$ -positive and  $x$ -negative directions, respectively. We may write displacements in the multilayered plate in the positive direction as

$$u_x = \sum_{n=0}^{\infty} \sum_{m=1}^4 A_{A+}^{(n)} d_m^{(n)} e^{ik\alpha_m^{(n)} z^{(n)}} e^{ik^{(n)}(x-c^{(n)}t)} \quad (33)$$

$$u_z = \sum_{n=0}^{\infty} \sum_{m=1}^4 A_{A+}^{(n)} d_m^{(n)} \beta_m^{(n)} e^{ik\alpha_m^{(n)} z^{(n)}} e^{ik^{(n)}(x-c^{(n)}t)} \quad (34)$$

where  $A_{A+}^{(n)}$  is the amplitude of mode  $n$  which will be determined and  $k^{(n)}$  denotes wavenumber. Parameters  $\alpha_m^{(n)}, \beta_m^{(n)}, d_m^{(n)}$  are depending on material properties and velocity of the  $n^{\text{th}}$  mode. Note that stress components can be easily calculated using Hooke's law.

We first choose state  $B$  to be a free guided wave of mode  $p$  in the negative direction. Substitution the expressions of state  $A$  and state  $B$  into Eq. (31), after manipulation, obtains

$$A_{A+}^{(p)} = - \frac{P U_z^{(p)(l)}(z_0)}{2I_A^{(p)}} \quad (35)$$

Here,  $p$  indicates the wave mode and  $l$  is the layer where the loading is applied. Similarly, with state  $B$  in the positive direction we find

$$A_{A-}^{(p)} = - \frac{P U_z^{(p)(l)}(z_0)}{2I_A^{(p)}} = A_{A+}^{(p)} \quad (36)$$

In Eqs. (35) and (36)

$$U_z^{p(l)} = \sum_{m=1}^4 d_m^{(p)} \beta_m^{(p)} e^{ik\alpha_m^{(p)} z^{(p)}} \quad (37)$$

For a horizontal load of magnitude  $Q$ , we have

$$A_{S+}^{(p)} = \frac{Q U_x^{(p)(l)}(z_0)}{2I_S^{(p)}}, \quad A_{S-}^{(p)} = - \frac{Q U_x^{(p)(l)}(z_0)}{2I_S^{(p)}} \quad (38)$$

For the case of multilayered structures, it is generally difficult to obtain closed-form expressions for integrals  $I_A^{(p)}$  and  $I_S^{(p)}$  so a numerical calculation is then considered.

## V. CONCLUSIONS

It has been discussed in this article the guided wave motions in a laminated composite plate subjected to time-harmonic sources. The dispersion relation of guided waves has been derived and the dispersion curves have been found. We have then obtained the closed-form expressions of guided waves in a composite plate under an ultrasonic load. The amplitudes of guided waves in positive and negative  $x$ -direction due to the load have derived.

## ACKNOWLEDGMENT

This research is funded by Graduate University of Science and Technology under grant number GUST.STS.DT2020-CH01.

## REFERENCES

- [1] V. Giurgiutiu, Structural Health Monitoring of Aerospace Composites, Elsevier Science, 2015.
- [2] J.L. Rose, Ultrasonic Guided Waves in Solid Media, Cambridge University Press, 2014.
- [3] S.K.S. Datta, A. H., Elastic Waves in Composite Media and Structures With Applications to Ultrasonic Nondestructive Evaluation, Taylor & Francis, 2009.
- [4] S. Rokhlin, S.R.D.C.P. Nagy, D. Chimenti, and P. Nagy, Physical Ultrasonics of Composites, Oxford University Press, 2011.
- [5] A.H. Nayfeh, and D.E. Chimenti, "Propagation of guided waves in fluid - coupled plates of fiber - reinforced composite", J. Acoust. Soc. Am., vol. 83 (5), pp. 1736-1743, 1988.
- [6] C.C. Habeger, R.W. Mann, and G.A. Baum, "Ultrasonic plate waves in paper", Ultrasonics, vol. 17 (2), pp. 57-62, 1979.
- [7] J.D. Achenbach, Wave Propagation in Elastic Solids, North-Holland Publishing Company, 1973.
- [8] J.D. Achenbach, Reciprocity in Elastodynamics, Cambridge University Press, 2003.

- [9] H. Phan, Y. Cho, and W. Li, "A theoretical approach to multiple scattering of surface waves by shallow cavities in a half-space", *Ultrasonics*, vol. 88, pp. 16-25, 2018.
- [10] H. Phan, Y. Cho, and J.D. Achenbach, "Application of the reciprocity theorem to scattering of surface waves by a cavity", *Int. J. Solids Struct.*, vol. 50 (24), pp. 4080-4088, 2013.
- [11] H. Phan, Y. Cho, and J.D. Achenbach, "Validity of the reciprocity approach for determination of surface wave motion", *Ultrasonics*, vol. 53 (3), pp. 665-671, 2013.
- [12] O. Balogun, and J.D. Achenbach, "Surface waves generated by a line load on a half-space with depth-dependent properties", *Wave Motion*, vol. 50 (7), pp. 1063-1072, 2013.
- [13] H. Phan, Y. Cho, Q.H. Le, C.V. Pham, H.T.L. Nguyen, P.T. Nguyen, and T.Q. Bui, "A closed-form solution to propagation of guided waves in a layered half-space under a time-harmonic load: An application of elastodynamic reciprocity", *Ultrasonics*, vol. 96, pp. 40-47, 2019.
- [14] P.T. Nguyen, H. Nguyen, D. Le, and H. Phan, "A model for ultrasonic guided waves in a cortical bone plate coupled with a soft-tissue layer", *AIP Conf. Proc.*, vol. 2102 (1), pp. 050007, 2019.
- [15] H. Phan, Y. Cho, C.V. Pham, H. Nguyen, and T.Q. Bui, "A theoretical approach for guided waves in layered structures", *AIP Conf. Proc.*, vol. 2102 (1), pp. 050011, 2019.
- [16] P.-T. Nguyen, and H. Phan, "A theoretical study on propagation of guided waves in a fluid layer overlying a solid half-space", *Vietnam Journal of Mechanics*, vol. 41 (1), pp. 51-62, 2019.
- [17] H. Phan, T.Q. Bui, H.T.L. Nguyen, and C.V. Pham, "Computation of interface wave motions by reciprocity considerations", *Wave Motion*, vol. 79, pp. 10-22, 2018.
- [18] D.K. Dao, V. Ngo, H. Phan, C.V. Pham, J. Lee, and T.Q. Bui, "Rayleigh wave motions in an orthotropic half-space under time-harmonic loadings: A theoretical study", *Applied Mathematical Modelling*, vol. 87, pp. 171-179, 2020.
- [19] H. Phan, Y. Cho, and J.D. Achenbach, "Verification of surface wave solutions obtained by the reciprocity theorem", *Ultrasonics*, vol. 54 (7), pp. 1891-1894, 2014.
- [20] A.H. Nayfeh, *Wave Propagation in Layered Anisotropic Media: With Applications to Composites*, Elsevier, 1995.

# Guided Wave Propagation in a Layered Half-Space Structure of Anisotropic Materials

Duy Kien Dao

Faculty of Civil Engineering  
Ho Chi Minh city of Technology and  
Education  
Ho Chi Minh City, Vietnam  
kiendd@hcmute.edu.vn

Ducho Le\*

Faculty of Mathematics, Mechanics  
and Informatics  
VNU University of Science  
Hanoi, Vietnam  
duchthole24@gmail.com

Quang Hung Le

Graduate University of Science and  
Technology  
VAST  
Hanoi, Vietnam  
lequanghungtdh@gmail.com

Minh Tuan Nguyen

Institute of Mechanics  
VAST  
Hanoi, Vietnam  
nmtuan@imech.vast.vn

Haidang Phan

Institute of Theoretical and Applied  
Research  
Duy Tan University  
Hanoi, Vietnam  
Faculty of Civil Engineering  
Duy Tan University  
Danang, Vietnam  
phanhaidang2@duytan.edu.vn

Duc Chinh Pham

Institute of Mechanics  
VAST  
Hanoi, Vietnam  
pdchinh@imech.vast.vn

**Abstract**—Layered half-spaces are widely used in engineering particularly for structures working under extreme conditions or having specific surface requirements. The thin layer helps improve engineering properties for the substrate material and prevent external damages caused by weather condition, friction, or chemical corrosion. Ultrasonic guided waves have been shown the advantages in material characterization and nondestructive evaluation of layered media that have a large dimension, include hard-to-reach areas, or are buried. In this paper, the motion of Rayleigh surface waves in layered half-spaces modeled as an orthotropic layer overlying an orthotropic half-space is investigated. The explicit expressions of free Rayleigh waves in both the layer and the half-space are first presented. Using the boundary conditions on the free surface and at the interface, the dispersive relations of Rayleigh waves are derived resulting in the dispersion curves. Reciprocity theorems are then applied to acquire the exact solutions of Rayleigh waves in coated structures of anisotropic materials under time-harmonic loads. The closed-form amplitudes of surface waves are found by using reciprocity theorems between two states, the Rayleigh wave that is generated by the time-harmonic forces and a free Rayleigh wave traveling in the layered structures.

**Keywords**—layered half-space, anisotropic material, Rayleigh waves, reciprocity theorem, time-harmonic source.

## I. INTRODUCTION

Over a few past decades, layered materials have drawn a great interest from numerous scholars because of its practical application in engineering and industry. A structure with one or more thin layers attached to a thick layer modelled as a layered half-space is one of them. With thin films, new mechanic, electronic, optic, magnetic and thermotic properties can be found and developed. Furthermore, the coated materials can efficiently counteract the external damaging factors including, for example, waterproof, anti-corrosion, mildew resistance and thermal barrier [1, 2]. Several machine components need to be coated to improve the tribological properties [3].

There may occur delamination at the interfaces between materials due to temperature change, external impact, or fatigue. Ultrasonic guided waves have demonstrated the ability and effectiveness to detect and evaluate this kind of defect [4]. Rayleigh waves propagate in an isotropic half-space was first investigated in [5] and later studied in, for example, [6]. Dispersive relation of free waves in a layer supporting by a semi-infinite space is described in [7]. Elastic surface waves guided by thin films is mathematically modeled and verified with experiment in [8]. For layered structures, excellent textbooks written by [4, 9, 10] were published. In the case of free waves, [11] studied Rayleigh and Love waves propagating in a coated half-space in which the isotropic thin layer was bonded to the transversely isotropic half-space. [12] showed the approximate secular equation and the formula of wave velocity in an orthotropic layered half-space. Different types of contacts between layer and half-space can be occurred corresponding to different kinds of boundary conditions. Smooth and sliding contacts were introduced in [13] and [14], respectively. Coated half-space with clamped surface was mentioned in [15]. Recent contributions on guided waves in layered half-spaces were reported in [16] in which the closed-form solutions of Rayleigh and Love wave motions under time-harmonic loads were obtained.

Reciprocity theorem was shown to be efficiently used in solving guided wave propagation and scattering problems in [17-20]. In a simple manner, the author computed the scattered amplitudes of surface waves and Lamb waves due to a time-harmonic load. The verification of the reciprocity approach was reported in [21-23]. The reciprocity relations were then applied to obtain closed-form solutions of guided waves in layered structures of isotropic materials [24-26].

In this current study, we introduce the explicit and compact expressions of Rayleigh waves in an anisotropic layered half-space. Afterward, the reciprocity theorem is employed to obtain the close-form solutions of Rayleigh wave motions generated by a time-harmonic source. With the aid of the explicit Rayleigh wave expressions, the complexity in computation due to the anisotropic nature of material are simplified.

\*Corresponding author: duchthole24@gmail.com

## II. EXPRESSIONS OF RAYLEIGH WAVES PROPAGATING IN AN ORTHOTROPIC LAYERED HALF-SPACE

Consider a half-space ( $\hat{\Omega}$ ) coated by a thin layer ( $\Omega$ ) of thickness  $h$  which have been shown in Fig. 1. The material of both the layer and the half-space which is of interest in the current work is elastic, homogeneous, and orthotropic. The elastic constants used for orthotropic material are indicated in stiffness matrix  $C$  and  $\hat{C}$  which corresponding to the properties of layer and half-space. The governing equation for elastic waves traveling in an orthotropic medium can be written as

$$\sigma_{ij,i} = \rho \ddot{u}_j \quad i, j = x, y, z \quad (1)$$

where  $\sigma_{ij}$  are stresses,  $u_i$  are displacements and,  $\rho$  denotes the volume mass density.

According to the superposition of partial wave method mentioned in [27], guided waves propagation in an orthotropic layer can be separated as six partial waves that reflect back ( $P^-$ ,  $SV^-$ ,  $SH^-$ ) and forth ( $P^+$ ,  $SV^+$ ,  $SH^+$ ) between the boundaries. However, there are only three partial waves exist in the half-space because it is infinite about  $z$ -direction; therefore, it has no reflection.

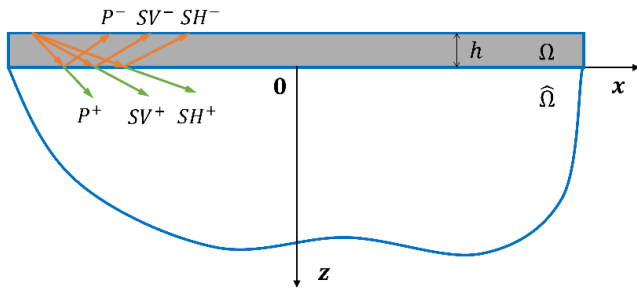


Fig. 1. Guided waves traveling in layered half-space structure

In this paper, we are concerned with the two-dimensional plane strain problem, i.e.,  $P$  and  $SV$  waves. The expressions of the relevant stresses can be expressed by Hooke's law for the orthotropic material as.

$$\sigma_{zz} = C_{13}u_{x,x} + C_{33}u_{z,z} \quad (2)$$

$$\sigma_{xz} = C_{55}(u_{x,z} + u_{z,x}) \quad (3)$$

$$\sigma_{xx} = C_{11}u_{x,x} + C_{13}u_{z,z} \quad (4)$$

The expressions of plane waves are contributed by four partial waves in the layer,

$$u_x = (A_1 e^{ik\alpha_1 z} + A_2 e^{ik\alpha_2 z} + A_3 e^{ik\alpha_3 z} + A_4 e^{ik\alpha_4 z}) e^{ik(x-ct)} \quad (5)$$

$$u_z = (W_1 A_1 e^{ik\alpha_1 z} + W_2 A_2 e^{ik\alpha_2 z} + W_3 A_3 e^{ik\alpha_3 z} + W_4 A_4 e^{ik\alpha_4 z}) e^{ik(x-ct)} \quad (6)$$

and two partial waves in the half-space, i.e.,

$$\hat{u}_x = (\hat{A}_1 e^{-k\hat{\alpha}_1 z} + \hat{A}_2 e^{-k\hat{\alpha}_2 z}) e^{ik(x-ct)} \quad (7)$$

$$\hat{u}_z = i(\hat{W}_1 \hat{A}_1 e^{-k\hat{\alpha}_1 z} + \hat{W}_2 \hat{A}_2 e^{-k\hat{\alpha}_2 z}) e^{ik(x-ct)} \quad (8)$$

where  $k$  is wavenumber in  $x$ -direction,  $A_j$  ( $j = \overline{1,4}$ ) and  $\hat{A}_j$  ( $j = 1,2$ ) are the unknown amplitudes of partial waves traveling in the layer and the half-space, respectively. These coefficients  $W_j$  and  $\hat{W}_j$  are the displacement ratios of  $z$  to  $x$ -direction, i.e.,

$$W_j = \frac{\rho c^2 - C_{11} - C_{55}\alpha_j^2}{(C_{13} + C_{55})\alpha_j} \quad (9)$$

$$\hat{W}_j = \frac{\hat{C}_{11} - \hat{\rho}c^2 - \hat{C}_{55}\hat{\alpha}_j^2}{(\hat{C}_{13} + \hat{C}_{55})\hat{\alpha}_j} \quad (10)$$

where  $\alpha_j$  and  $\hat{\alpha}_j$  are the ratios of the wavenumbers between thickness  $z$ -direction and  $x$ -direction. They can be found by solving Christoffel equation as

$$\alpha_1 = -\alpha_3 = \sqrt{\frac{S + \sqrt{S^2 - 4P}}{2}} \quad (11)$$

$$\alpha_2 = -\alpha_4 = \sqrt{\frac{S - \sqrt{S^2 - 4P}}{2}}$$

$$\hat{\alpha}_1 = \sqrt{\frac{\hat{S} + \sqrt{\hat{S}^2 - 4\hat{P}}}{2}} ; \quad \hat{\alpha}_2 = \sqrt{\frac{\hat{S} - \sqrt{\hat{S}^2 - 4\hat{P}}}{2}} \quad (12)$$

where

$$S = \frac{(c_{13} + c_{55})^2 - (c_{11} - \rho c^2)c_{33} - (c_{55} - \rho c^2)c_{55}}{c_{33}c_{55}} \quad (13)$$

$$P = \frac{(c_{11} - \rho c^2)(c_{55} - \rho c^2)}{c_{33}c_{55}} \quad (14)$$

$$\hat{S} = \frac{(\hat{c}_{11} - \hat{\rho}c^2)\hat{c}_{33} + (\hat{c}_{55} - \hat{\rho}c^2)\hat{c}_{55} - (\hat{c}_{13} + \hat{c}_{55})^2}{\hat{c}_{33}\hat{c}_{55}} \quad (15)$$

$$\hat{P} = \frac{(\hat{c}_{11} - \hat{\rho}c^2)(\hat{c}_{55} - \hat{\rho}c^2)}{\hat{c}_{33}\hat{c}_{55}} \quad (16)$$

Substitute displacements into Eqs. (2) – (4) yield,



$$\sigma_{zz} = \sum_{j=1}^4 ikD_{1j}A_j e^{ik\alpha_j z} e^{ik(x-ct)} \quad (17)$$

$$\sigma_{xz} = \sum_{j=1}^4 ikD_{2j}A_j e^{ik\alpha_j z} e^{ik(x-ct)} \quad (18)$$

$$\sigma_{xx} = \sum_{j=1}^4 ikD_{3j}A_j e^{ik\alpha_j z} e^{ik(x-ct)} \quad (19)$$

where

$$D_{1j} = C_{13} + C_{33}\alpha_j W_j \quad (20)$$

$$D_{2j} = C_{55}(\alpha_j + W_j) \quad (21)$$

$$D_{3j} = C_{11} + C_{13}\alpha_j W_j \quad (22)$$

Similarity, the expressions of stresses in half-space can be written as

$$\hat{\sigma}_{zz} = ik\hat{D}_{1j}(\hat{A}_1 e^{-k\hat{\alpha}_1 z} + \hat{A}_2 e^{-k\hat{\alpha}_2 z}) e^{ik(x-ct)} \quad (23)$$

$$\hat{\sigma}_{xz} = -k\hat{D}_{2j}(\hat{A}_1 e^{-k\hat{\alpha}_1 z} + \hat{A}_2 e^{-k\hat{\alpha}_2 z}) e^{ik(x-ct)} \quad (24)$$

$$\hat{\sigma}_{xx} = ik\hat{D}_{3j}(\hat{A}_1 e^{-k\hat{\alpha}_1 z} + \hat{A}_2 e^{-k\hat{\alpha}_2 z}) e^{ik(x-ct)} \quad (25)$$

Where

$$\hat{D}_{1j} = \hat{C}_{13} - \hat{C}_{33}\hat{\alpha}_j \hat{W}_j \quad (26)$$

$$\hat{D}_{2j} = \hat{C}_{55}(\hat{\alpha}_j + \hat{W}_j) \quad (27)$$

$$\hat{D}_{3j} = \hat{C}_{11} - \hat{C}_{13}\hat{\alpha}_j \hat{W}_j \quad (28)$$

Here, we are interested in the perfectly bonded interface between layer and half-space. This property leads to the boundary conditions at the interface which indicate the continuous in displacements and the balance in stresses. Combined with the free conditions at the top of layer we obtain the boundary conditions of entire structure, i.e.,

$$\sigma_{zz} = 0; \quad \sigma_{xz} = 0 \quad \text{at } z = -h \quad (29)$$

$$u_x = \hat{u}_x; \quad u_z = \hat{u}_z; \quad \sigma_{zz} = \hat{\sigma}_{zz}; \quad \sigma_{xz} = \hat{\sigma}_{xz} \quad \text{at } z = 0 \quad (30)$$

Substitute the displacements and stresses into the boundary conditions yields,

$$ik(D_{11}A_1 e^{-ik\alpha_1 h} + D_{12}A_2 e^{-ik\alpha_2 h} + D_{11}A_3 e^{ik\alpha_1 h} + D_{12}A_4 e^{ik\alpha_2 h}) e^{ik(x-ct)} = 0 \quad (31)$$

$$ik(D_{21}A_1 e^{-ik\alpha_1 h} + D_{22}A_2 e^{-ik\alpha_2 h} - D_{21}A_3 e^{ik\alpha_1 h} - D_{22}A_4 e^{ik\alpha_2 h}) e^{ik(x-ct)} = 0 \quad (32)$$

$$(A_1 + A_2 + A_3 + A_4 - \hat{A}_1 - \hat{A}_2) e^{ik(x-ct)} = 0 \quad (33)$$

$$(W_1 A_1 + W_2 A_2 - W_1 A_3 - W_2 A_4 - i\hat{W}_1 \hat{A}_1 - i\hat{W}_2 \hat{A}_2) e^{ik(x-ct)} = 0 \quad (34)$$

$$ik(D_{11}A_1 + D_{12}A_2 + D_{11}A_3 + D_{12}A_4 - \hat{D}_{11}\hat{A}_1 - \hat{D}_{12}\hat{A}_2) e^{ik(x-ct)} = 0 \quad (35)$$

$$ik(D_{21}A_1 + D_{22}A_2 - D_{21}A_3 - D_{22}A_4 - i\hat{D}_{21}\hat{A}_1 - i\hat{D}_{22}\hat{A}_2) e^{ik(x-ct)} = 0 \quad (36)$$

Equations (31)–(36) define a set of six homogenous linear equations for displacement amplitudes  $A_j$  and  $\hat{A}_j$ . These equations can be rewritten in matrix notation as,

$$\mathbf{T}\mathbf{A} = 0 \quad (37)$$

where

$$\mathbf{A} = [A_1 \ A_2 \ A_3 \ A_4 \ \hat{A}_1 \ \hat{A}_2]^T \quad (38)$$

$$\mathbf{T} = \begin{bmatrix} D_{11}e^{-ik\alpha_1 h} & D_{12}e^{-ik\alpha_2 h} & D_{11}e^{ik\alpha_1 h} & D_{12}e^{ik\alpha_2 h} & 0 & 0 \\ D_{21}e^{-ik\alpha_1 h} & D_{22}e^{-ik\alpha_2 h} & -D_{21}e^{ik\alpha_1 h} & -D_{22}e^{ik\alpha_2 h} & 0 & 0 \\ 1 & 1 & 1 & 1 & -1 & -1 \\ W_1 & W_2 & -W_1 & -W_2 & -i\hat{W}_1 & -i\hat{W}_2 \\ D_{11} & D_{12} & D_{11} & D_{12} & -\hat{D}_{11} & -\hat{D}_{12} \\ D_{21} & D_{22} & -D_{21} & -D_{22} & -i\hat{D}_{21} & -i\hat{D}_{22} \end{bmatrix} \quad (39)$$

For nontrivial solutions, the determinant of matrix  $\mathbf{T}$  in Eq. (39) must be zero, i.e.,

$$\det(\mathbf{T}) = 0 \quad (40)$$

In Eq. (40), phase velocity and frequency are unknown variable while other coefficients depend on material properties of the layer and the half-space. Therefore, we always find at least one value of phase velocity corresponding to a given frequency. The dependent relation between phase velocity and frequency is called dispersive relation and Eq. (40) is called characteristic equation. By solving the characteristic equation, the dispersive relation of Rayleigh waves will result in the dispersion curves. Subsequently, a numerical procedure depended on the interpolation method is employed to plot the dispersion curves.

As an example, we consider an orthotropic layered half-space with material properties of the layer and the half-space are given in. The dispersion curves are shown in TABLE 1. The dispersion curves are shown in Fig. 2.

TABLE 1. MATERIAL PROPERTIES OF LAYER AND HALF-SPACE

Material	Stiffness matrix (GPa)	Density (kg/m <sup>3</sup> )
Graphite/ Epoxy (layer)	$C = \begin{bmatrix} 134.9 & 5.2 & 5.2 & 0 & 0 & 0 \\ 5.2 & 14.4 & 7.1 & 0 & 0 & 0 \\ 5.2 & 7.1 & 14.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5.6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5.6 \end{bmatrix}$	1800
Aluminum (half-space)	$C = \begin{bmatrix} 143.8 & 6.2 & 6.2 & 0 & 0 & 0 \\ 6.2 & 13.3 & 6.5 & 0 & 0 & 0 \\ 6.2 & 6.5 & 13.3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5.7 \end{bmatrix}$	2700

The thickness of layer is chosen as  $h = 1$  mm. Both phase velocity and group velocity of Rayleigh waves are illustrated, i.e.,

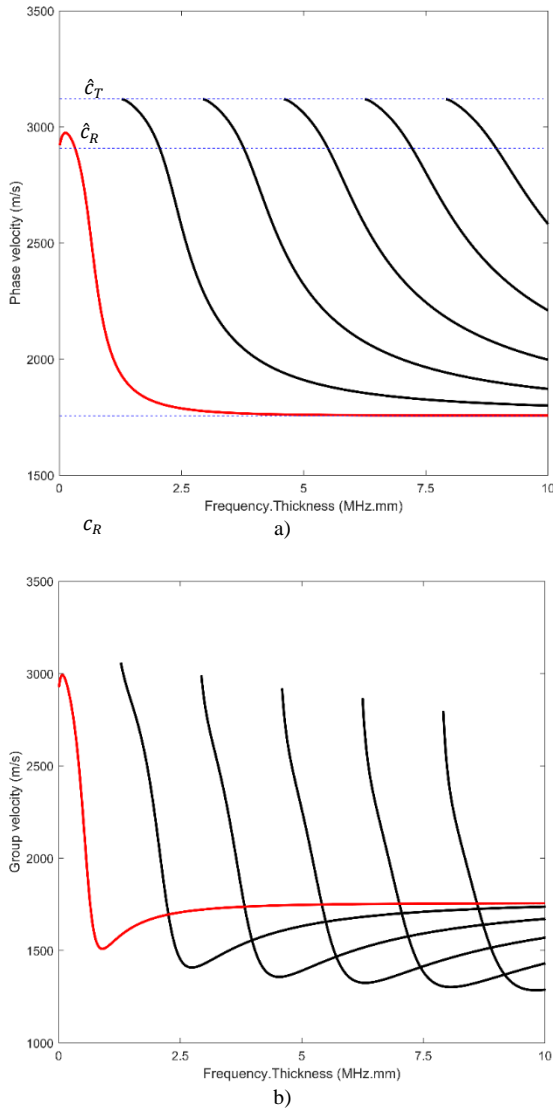


Fig. 2. Dispersion curve of Rayleigh waves in an orthotropic layered half-space. a) Phase velocity; b) Group velocity

It is conspicuous that the phase velocity of the first Rayleigh mode approaches the value of velocity of Rayleigh wave in the half-space ( $\hat{c}_R$ ) when the thickness of layer comes to zero. The upper limit of dispersion curves is equal to the transvers velocity of Rayleigh wave in the half-space ( $\hat{c}_T$ ). This threshold velocity is the result of the existing condition of Rayleigh waves [14]. The lower limit of dispersion curves is the Rayleigh wave velocity of the layer ( $c_R$ ). All dispersion curves are asymptotic to the lower limit line when the quantity  $fh$  increases to infinity.

There are six unknown parameters  $A_1, A_2, A_3, A_4, \hat{A}_1$  and  $\hat{A}_2$  containing in Eq. (37) while it only has five dependent equations. Therefore, Eq. (37) has an infinite number of solutions. For the sake of simplicity, general solutions of Eq. (37) are written in form

$$A_1 = Ad_1; A_2 = Ad_2; A_3 = Ad_3; A_4 = Ad_4; \quad (41)$$

$$\hat{A}_1 = A\hat{d}_1; \hat{A}_2 = A\hat{d}_2$$

where  $A$  is an arbitrary constant with dimension of length while coefficients  $d_1, d_2, d_3, d_4, \hat{d}_1$  and  $\hat{d}_2$  are dimensionless and only depend on material properties. After some manipulation, these coefficients can be derived as,

$$d_1 = H_{11}\hat{d}_1 + H_{12}\hat{d}_2 \quad (42)$$

$$d_2 = H_{21}\hat{d}_1 + H_{22}\hat{d}_2 \quad (43)$$

$$d_3 = H_{31}\hat{d}_1 + H_{32}\hat{d}_2 \quad (44)$$

$$d_4 = H_{41}\hat{d}_1 + H_{42}\hat{d}_2 \quad (45)$$

$$\hat{d}_1 = \frac{D_{11}}{s_1}(H_{12} + s_1^2 H_{32}) + \frac{D_{12}}{s_2}(H_{22} + s_2^2 H_{42}) \quad (46)$$

$$\hat{d}_2 = -\frac{D_{11}}{s_1}(H_{11} + s_1^2 H_{31}) - \frac{D_{12}}{s_2}(H_{21} + s_2^2 H_{41}) \quad (47)$$

where

$$s_1 = e^{ik\alpha_1 h}; \quad s_2 = e^{ik\alpha_2 h} \quad (48)$$

$$H_{11} = \frac{\hat{D}_{11} - D_{12}}{2(D_{11} - D_{12})} + i \frac{\hat{D}_{21}W_2 - D_{22}\hat{W}_1}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{12} = \frac{\hat{D}_{12} - D_{12}}{2(D_{11} - D_{12})} + i \frac{\hat{D}_{22}W_2 - D_{22}\hat{W}_2}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{21} = \frac{D_{11} - \hat{D}_{11}}{2(D_{11} - D_{12})} + i \frac{D_{21}\hat{W}_1 - \hat{D}_{21}W_1}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{22} = \frac{D_{11} - \widehat{D}_{12}}{2(D_{11} - D_{12})} + i \frac{D_{21}\widehat{W}_2 - \widehat{D}_{22}W_1}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{31} = \frac{\widehat{D}_{11} - D_{12}}{2(D_{11} - D_{12})} - i \frac{\widehat{D}_{21}W_2 - D_{22}\widehat{W}_1}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{32} = \frac{\widehat{D}_{12} - D_{12}}{2(D_{11} - D_{12})} - i \frac{\widehat{D}_{22}W_2 - D_{22}\widehat{W}_2}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{41} = \frac{D_{11} - \widehat{D}_{11}}{2(D_{11} - D_{12})} - i \frac{D_{21}\widehat{W}_1 - \widehat{D}_{21}W_1}{2(D_{21}W_2 - D_{22}W_1)}$$

$$H_{42} = \frac{D_{11} - \widehat{D}_{12}}{2(D_{11} - D_{12})} - i \frac{D_{21}\widehat{W}_2 - \widehat{D}_{22}W_1}{2(D_{21}W_2 - D_{22}W_1)}$$

Consequently, the expressions of displacements and stresses in layer can be rewritten in a more compact form, i.e.,

$$u_x = AU_x(z)e^{ik(x-ct)} \quad (49)$$

$$u_z = AU_z(z)e^{ik(x-ct)} \quad (50)$$

$$\sigma_{xx} = ikC_{55}AT_{xx}(z)e^{ik(x-ct)} \quad (51)$$

$$\sigma_{xz} = ikC_{55}AT_{xz}(z)e^{ik(x-ct)} \quad (52)$$

where

$$U_x(z) = d_1e^{ik\alpha_1z} + d_2e^{ik\alpha_2z} + d_3e^{-ik\alpha_1z} + d_4e^{-ik\alpha_2z} \quad (53)$$

$$U_z(z) = W_1d_1e^{ik\alpha_1z} + W_2d_2e^{ik\alpha_2z} - W_1d_3e^{-ik\alpha_1z} - W_2d_4e^{-ik\alpha_2z} \quad (54)$$

$$T_{xx}(z) = \frac{D_{31}}{C_{55}}d_1e^{ik\alpha_1z} + \frac{D_{32}}{C_{55}}d_2e^{ik\alpha_2z} + \frac{D_{31}}{C_{55}}d_3e^{-ik\alpha_1z} + \frac{D_{32}}{C_{55}}d_4e^{-ik\alpha_2z} \quad (55)$$

$$T_{xz}(z) = (\alpha_1 + W_1)d_1e^{ik\alpha_1z} + (\alpha_2 + W_2)d_2e^{ik\alpha_2z} - (\alpha_1 + W_1)d_3e^{-ik\alpha_1z} - (\alpha_2 + W_2)d_4e^{-ik\alpha_2z} \quad (56)$$

In orthotropic half-space, we also have similar expressions, i.e.,

$$\hat{u}_x = A\hat{U}_x(z)e^{ik(x-ct)} \quad (57)$$

$$\hat{u}_z = iA\hat{U}_z(z)e^{ik(x-ct)} \quad (58)$$

$$\hat{\sigma}_{xx} = ik\hat{C}_{55}A\hat{T}_{xx}(z)e^{ik(x-ct)} \quad (59)$$

$$\hat{\sigma}_{xz} = -k\hat{C}_{55}A\hat{T}_{xz}(z)e^{ik(x-ct)} \quad (60)$$

where

$$\hat{U}_x(z) = \hat{d}_1e^{-k\hat{\alpha}_1z} + \hat{d}_2e^{-k\hat{\alpha}_2z} \quad (61)$$

$$\hat{U}_z(z) = \hat{W}_1\hat{d}_1e^{-k\hat{\alpha}_1z} + \hat{W}_2\hat{d}_2e^{-k\hat{\alpha}_2z} \quad (62)$$

$$\hat{T}_{xx}(z) = \frac{\widehat{D}_{31}}{\hat{C}_{55}}\hat{d}_1e^{-k\hat{\alpha}_1z} + \frac{\widehat{D}_{32}}{\hat{C}_{55}}\hat{d}_2e^{-k\hat{\alpha}_2z} \quad (63)$$

$$\hat{T}_{xz}(z) = (\hat{\alpha}_1 + \hat{W}_1)\hat{d}_1e^{-k\hat{\alpha}_1z} + (\hat{\alpha}_2 + \hat{W}_2)\hat{d}_2e^{-k\hat{\alpha}_2z} \quad (64)$$

Here, these coefficients  $U_x, U_z, T_{xx}, T_{xz}$  and  $\hat{U}_x, \hat{U}_z, \hat{T}_{xx}, \hat{T}_{xz}$  are functions of depth  $z$ . By writing displacements and stresses in this form, it only has one unknown variable  $A$  need to be solved. Consequently, it is convenient to compute amplitude of Rayleigh waves by reciprocity theorems which will be discussed in the following section.

### III. COMPUTATION OF RAYLEIGH WAVES MOTION USING RECIPROCITY THEOREMS.

Reciprocity theorems can be used to obtain solutions of actual field conjunction with a fundamental solution of free waves which called virtual waves. In current problem, actual state is Rayleigh wave motions generated by a time-harmonic load in orthotropic layered half-space structure. The formula of reciprocity theorems for two-material body can be written as a combination of two domain integrals, see [17].

$$\begin{aligned} & \int_{\Omega} (f_j^A u_j^A - f_j^B u_j^B) d\Omega + \int_{\hat{\Omega}} (\hat{f}_j^A \hat{u}_j^A - \hat{f}_j^B \hat{u}_j^B) d\hat{\Omega} \\ &= \int_S (\sigma_{ij}^B u_j^A - \sigma_{ij}^A u_j^B) n_i dS \\ &+ \int_{\hat{S}} (\hat{\sigma}_{ij}^B \hat{u}_j^A - \hat{\sigma}_{ij}^A \hat{u}_j^B) \hat{n}_i d\hat{S} \end{aligned} \quad (65)$$

Here,  $S$  and  $\hat{S}$  are the boundaries of layer and half-space, respectively, while  $n$  and  $\hat{n}$  are normal vectors which have been shown in Fig. 3. The superscripts <sup>A</sup> and <sup>B</sup> denote the actual and virtual states, respectively.

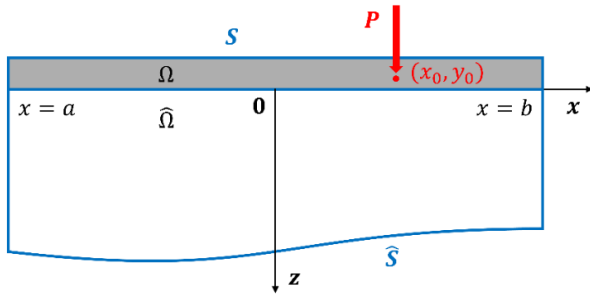


Fig. 3. Layered half-space under a time-harmonic source

We first consider a vertical load which applied at  $(x_0, z_0)$  point in form of Delta function, i.e.,

$$f_z^A = P\delta(z - z_0)\delta(x - x_0)e^{-ikct} \quad (66)$$

Under time-harmonic load, both body waves and Rayleigh waves are generated but in the far field, only Rayleigh waves are existed because of the attenuation in exponential function of body waves, see [18]. The motions of Rayleigh waves can be expressed as a summation of wave modes, see [16].

For actual state A, the expressions of displacements and stresses in the far field at the positive  $x$  – direction can be written in layer as

$$u_x^m = \sum_{m=0}^{\infty} A_m^{P+} U_x^m(z) e^{ik_m(x-c_mt)} \quad (67)$$

$$u_z^m = \sum_{m=0}^{\infty} A_m^{P+} U_z^m(z) e^{ik_m(x-c_mt)} \quad (68)$$

$$\sigma_{xx}^m = \sum_{m=0}^{\infty} ik_m C_{55} A_m^{P+} T_{xx}^m(z) e^{ik_m(x-c_mt)} \quad (69)$$

$$\sigma_{xz}^m = \sum_{m=0}^{\infty} ik_m C_{55} A_m^{P+} T_{xz}^m(z) e^{ik_m(x-c_mt)} \quad (70)$$

and in orthotropic half-space as

$$\hat{u}_x^m = \sum_{m=0}^{\infty} A_m^{P+} \hat{U}_x^m(z) e^{ik_m(x-c_mt)} \quad (71)$$

$$\hat{u}_z^m = \sum_{m=0}^{\infty} iA_m^{P+} \hat{U}_z^m(z) e^{ik_m(x-c_mt)} \quad (72)$$

$$\hat{\sigma}_{xx}^m = \sum_{m=0}^{\infty} ik_m \hat{C}_{55} A_m^{P+} \hat{T}_{xx}^m(z) e^{ik_m(x-c_mt)} \quad (73)$$

$$\hat{\sigma}_{xz}^m = \sum_{m=0}^{\infty} -k_m \hat{C}_{55} A_m^{P+} \hat{T}_{xz}^m(z) e^{ik_m(x-c_mt)} \quad (74)$$

where  $A_m^{P+}$  and  $k_m$  are the amplitude and wavenumber of mode  $m$ , respectively, while  $U_x^m, U_z^m, T_{xx}^m, T_{xz}^m$  and  $\hat{U}_x^m, \hat{U}_z^m, \hat{T}_{xx}^m, \hat{T}_{xz}^m$  are the functions which have been defined in previous section.

In virtual state B, the Rayleigh mode  $n$  traveling in negative  $x$  – direction is chosen as

$$u_x^n = -B_n U_x^n(z) e^{-ik_n(x+c_nt)} \quad (75)$$

$$u_z^n = B_n U_z^n(z) e^{-ik_n(x+c_nt)} \quad (76)$$

$$\sigma_{xx}^n = ik_n C_{55} B_n T_{xx}^n(z) e^{-ik_n(x+c_nt)} \quad (77)$$

$$\sigma_{xz}^n = -ik_n C_{55} B_n T_{xz}^n(z) e^{-ik_n(x+c_nt)} \quad (78)$$

and

$$\hat{u}_x^n = -B_n \hat{U}_x^n(z) e^{-ik_n(x+c_nt)} \quad (79)$$

$$\hat{u}_z^n = -iB_n \hat{U}_z^n(z) e^{-ik_n(x+c_nt)} \quad (80)$$

$$\hat{\sigma}_{xx}^n = ik_n \hat{C}_{55} B_n \hat{T}_{xx}^n(z) e^{-ik_n(x+c_nt)} \quad (81)$$

$$\hat{\sigma}_{xz}^n = -ik_n \hat{C}_{55} B_n \hat{T}_{xz}^n(z) e^{-ik_n(x+c_nt)} \quad (82)$$

Follow the computation procedure in [16], the amplitude of Rayleigh waves can be obtained after substituting the displacements and stresses of two states into reciprocity formula in Eq. (65). Here, for the brief of paper, we introduce the final expression of manipulations, i.e.,

$$P U_z^n(z_0) e^{-ik_n x_0} = \sum_{m=0}^{\infty} A_m^{P+} e^{i(k_m - k_n)b} (C_{55} I_{mn} + \hat{C}_{55} \hat{I}_{mn}) \quad (83)$$

where

$$I_{mn} = \int_{-h}^0 i \left[ k_n (T_{xx}^n(z) U_x^m(z) - T_{xz}^n(z) U_z^m(z)) + k_m (T_{xx}^m(z) U_x^n(z) - T_{xz}^m(z) U_z^n(z)) \right] dz \quad (84)$$

$$\hat{I}_{mn} = \int_0^{\infty} i \left[ k_n (\hat{T}_{xx}^n(z) \hat{U}_x^m(z) - \hat{T}_{xz}^n(z) \hat{U}_z^m(z)) + k_m (\hat{T}_{xx}^m(z) \hat{U}_x^n(z) - \hat{T}_{xz}^m(z) \hat{U}_z^n(z)) \right] dz \quad (85)$$

These integrals  $I_{mn}$  and  $\hat{I}_{mn}$  only yield in the case with  $m = n$  because of the orthogonality relation, see [16], then we have the expression of amplitude

$$A_m^{P+} = \frac{PU_z^n(z_0)e^{-ik_n x_0}}{C_{55}I_{nn} + \hat{C}_{55}\hat{I}_{nn}} \quad (86)$$

where

$$I_{nn} = \int_{-h}^0 2ik_n (T_{xx}^n(z)U_x^n(z) - T_{xz}^n(z)U_z^n(z)) dz \quad (87)$$

$$\hat{I}_{nn} = \int_0^\infty 2ik_n (\hat{T}_{xx}^n(z)\hat{U}_x^n(z) - \hat{T}_{xz}^n(z)\hat{U}_z^n(z)) dz \quad (88)$$

If we choose the propagating direction of state  $B$  is positive  $x$  – direction, the amplitude is easily obtained by similar derivation as

$$A_m^{P-} = \frac{PU_z^n(z_0)e^{ik_n x_0}}{C_{55}I_{nn} + \hat{C}_{55}\hat{I}_{nn}} \quad (89)$$

Similar results for the case of horizontal load can be obtained straightforwardly.

#### IV. CONCLUSIONS

In this paper, we have studied on the propagation of Rayleigh waves in an orthotropic layered half-space in which the contact between layer and half-space is assumed that perfectly bonded. The explicit and compact expressions of Rayleigh waves propagating in orthotropic layered half-space are introduced. By using this explicit form, the Rayleigh wave motions in orthotropic half-space can be obtained more conveniently and straightforwardly. The dispersion curves of Rayleigh waves in orthotropic layered half-space are proposed. Depending on the behavior of lowest Rayleigh mode we can verify the properties of guided wave in layer and half-space. Reciprocity theorems are also employed to compute the displacement amplitude of Rayleigh waves generated by a time-harmonic source.

#### ACKNOWLEDGMENT

This research has been supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant reference 107.02-2019.21 and Graduate University of Science and Technology under grant number GUST.STS.ĐT2020-CH01.

#### REFERENCES

- [1] N.P. Padture, M. Gell, and E.H. Jordan, “Thermal Barrier Coatings for Gas-Turbine Engine Applications”, *Science*, vol. 296 (5566), pp. 280, 2002.
- [2] D.K. Chattopadhyay, and K.V.S.N. Raju, “Structural engineering of polyurethane coatings for high performance applications”, *Progress in Polymer Science*, vol. 32 (3), pp. 352-418, 2007.
- [3] S. Hogmark, S. Jacobson, and M. Larsson, “Design and evaluation of tribological coatings”, *Wear*, vol. 246 (1), pp. 20-33, 2000.
- [4] J.L. Rose, *Ultrasonic Guided Waves in Solid Media*, Cambridge University Press, 2014.
- [5] L. Rayleigh, “On Waves Propagated along the Plane Surface of an Elastic Solid”, *Proceedings of the London Mathematical Society*, vol. s1-17 (1), pp. 4-11, 1885.
- [6] J. Achenbach, *Wave propagation in elastic solids*, Elsevier, 1973.
- [7] J.D. Achenbach, and S.P. Keshava, “Free Waves in a Plate Supported by a Semi-Infinite Continuum”, *Journal of Applied Mechanics*, vol. 34 (2), pp. 397-404, 1967.
- [8] H.F. Tiersten, “Elastic Surface Waves Guided by Thin Films”, *Journal of Applied Physics*, vol. 40 (2), pp. 770-789, 1969.
- [9] A.H. Nayfeh, *Wave propagation in layered anisotropic media: With application to composites*, Elsevier, 1995.
- [10] V. Giurgiutiu, *Structural health monitoring of aerospace composites*, Academic Press, 2015.
- [11] M. Bouden, and S.K. Datta, “Rayleigh and Love Waves in Cladded Anisotropic Medium”, *Journal of Applied Mechanics*, vol. 57 (2), pp. 398-403, 1990.
- [12] P.C. Vinh, and N.T.K. Linh, “An approximate secular equation of Rayleigh waves propagating in an orthotropic elastic half-space coated by a thin orthotropic elastic layer”, *Wave Motion*, vol. 49 (7), pp. 681-689, 2012.
- [13] P.C. Vinh, V.T.N. Anh, and V.P. Thanh, “Rayleigh waves in an isotropic elastic half-space coated by a thin isotropic elastic layer with smooth contact”, *Wave Motion*, vol. 51 (3), pp. 496-504, 2014.
- [14] P.C. Vinh, and V.T. Ngoc Anh, “Rayleigh waves in an orthotropic half-space coated by a thin orthotropic layer with sliding contact”, *International Journal of Engineering Science*, vol. 75, pp. 154-164, 2014.
- [15] J. Kaplunov, D. Prikazchikov, and L. Sultanova, “Rayleigh-type waves on a coated elastic half-space with a clamped surface”, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 377 (2156), pp. 20190111, 2019.
- [16] H. Phan, Y. Cho, Q.H. Le, C.V. Pham, H.T.L. Nguyen, P.T. Nguyen, and T.Q. Bui, “A closed-form solution to propagation of guided waves in a layered half-space under a time-harmonic load: An application of elastodynamic reciprocity”, *Ultrasonics*, vol. 96, pp. 40-47, 2019.
- [17] J. Achenbach, *Reciprocity in elastodynamics*, Cambridge University Press, 2003.
- [18] H. Phan, Y. Cho, and J.D. Achenbach, “Application of the reciprocity theorem to scattering of surface waves by a cavity”, *International Journal of Solids and Structures*, vol. 50 (24), pp. 4080-4088, 2013.
- [19] H. Phan, Y. Cho, and W. Li, “A theoretical approach to multiple scattering of surface waves by shallow cavities in a half-space”, *Ultrasonics*, vol. 88, pp. 16-25, 2018.
- [20] P.H. Dang, L.D. Tho, L.Q. Hung, and D.D. Kien, “Investigation of Rayleigh wave interaction with surface defects”, *Journal of Science and Technology in Civil Engineering*, vol. 13 (3), pp. 95-103, 2019.
- [21] H. Phan, Y. Cho, and J.D. Achenbach, “Validity of the reciprocity approach for determination of surface wave motion”, *Ultrasonics*, vol. 53 (3), pp. 665-671, 2013.
- [22] H. Phan, Y. Cho, and J.D. Achenbach, “Verification of surface wave solutions obtained by the reciprocity theorem”, *Ultrasonics*, vol. 54 (7), pp. 1891-1894, 2014.
- [23] D.K. Dao, V. Ngo, H. Phan, C.V. Pham, J. Lee, and T.Q. Bui, “Rayleigh wave motions in an orthotropic half-space under time-harmonic loadings: A theoretical study”, *Applied Mathematical Modelling*, vol. 87, pp. 171-179, 2020.
- [24] H. Phan, Y. Cho, C.V. Pham, H. Nguyen, and T.Q. Bui, “A theoretical approach for guided waves in layered structures”, *AIP Conference Proceedings*, vol. 2102 (1), pp. 050011, 2019.
- [25] H. Phan, T.Q. Bui, H.T.L. Nguyen, and C.V. Pham, “Computation of interface wave motions by reciprocity considerations”, *Wave Motion*, vol. 79, pp. 10-22, 2018.
- [26] P.-T. Nguyen, and H. Phan, “A theoretical study on propagation of guided waves in a fluid layer overlying a solid half-space”, *Vietnam Journal of Mechanics*, vol. 41 (1), pp. 51-62, 2019.
- [27] L.P. Solie, and B.A. Auld, “Elastic waves in free anisotropic plates”, *The Journal of the Acoustical Society of America*, vol. 54 (1), pp. 50-65, 1973.



# City-Scale Electricity Demand Forecasting using a Gaussian Process Model

Phong T. T. Nguyen

Faculty of Civil Engineering  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
(formerly, Graduate Research Assistant,  
The University of Texas at Austin)  
Email: phongntt@hcmute.edu.vn

Lance Manuel

Department of Civil, Architectural  
and Environmental Engineering  
The University of Texas at Austin  
Austin, Texas, USA  
Email: lmanuel@utexas.edu

**Abstract**—Accurate and efficient power demand forecasting in urban settings is essential for making decisions related to planning, managing and operations in electricity supply. This task, however, is complicated due to many sources of uncertainty such as due to the variation in weather conditions and household or other needs that influence the inherent stochastic and nonlinear characteristics of electricity demand. Due to the modeling flexibility and computational efficiency afforded by it, a Gaussian process model is employed in this study for energy demand prediction as a function of temperature. A Gaussian process model is a Bayesian non-parametric regression method that models data using a joint Gaussian distribution with mean and covariance functions. The selected mean function is modeled as a polynomial function of temperature, whereas the covariance function is appropriately selected to reflect the actual data patterns. We employ real data sets of daily temperature and electricity demand from Austin, Texas, USA to assess the effectiveness of the proposed method for load forecasting. The accuracy of the model prediction is evaluated using metrics such as mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE) and 95% confidence interval (95% CI). A numerical study undertaken demonstrates that the proposed method has promise for energy demand prediction.

**Index Terms**—Gaussian process; electricity load forecasting, probability density forecasting.

## I. INTRODUCTION

Demand forecasting plays an important role for utility companies in the electricity industry, as it provides a basis for making decisions in power system planning, managing and operation. Accurate energy demand forecasting can improve the efficiency of power stations and ensure the safety of the grid. It is suggested that even a few percentage points increase in prediction accuracy can have a significant cost impact on companies operating in highly competitive power markets [1].

Various methods for predicting electricity demand have been considered in recent years [2]. Traditionally, time series methods such as ARIMA [3], [4], [5] have been often used for short-term prediction. Regression-based models [6], [7] are also simple and easy to implement. Such forecasting models, however, can lack desired accuracy in forecasting due to their limited capability of accounting for nonlinear characteristics [8].

Considering this complexity and potentially nonlinear nature as well as significant fluctuations with time of power loads, kernel-based methods have been proposed to deal with this forecasting or prediction problem. One such kernel-based method, involving the use of a Gaussian process [9] model, has become one of the more promising approaches due to its flexibility and good performance with empirical data in practical applications [10]. A key advantage of the use of a Gaussian process that allows for great modeling flexibility results from its non-parametric characteristics and great computational efficiency in both load forecasting and in uncertainty quantification through Bayesian updating. Additionally, various kernel functions, that are needed, are available for accurately modeling of complex patterns seen in historical data. Exploiting these advantages, other research studies have been conducted that have applied Gaussian process models for load forecasting and these have shown promise [10], [11], [12]. Other advanced machine learning techniques for loads forecasting continue to be explored in the literature [12], [13], [14], [15], [16]. Continued investigation and development of these methods is necessary to facilitate their practical application—it is important that all methods address minimal data requirements in training, quantification of the uncertainty due to uncertain inputs, and field validation across different cases.

Recognizing the strong dependence of utility loads on temperature noted in other studies [17], the present study implements a Gaussian process model for demand forecasting that utilizes external temperature as the only input and relies on historically observed energy consumption patterns to improve the accuracy and robustness of load predictions, while seeking to reduce the need for extensive training data. We also suggest a scenario-based framework for dealing with the uncertainty in load forecasting when the actual temperatures at the time of prediction are not available. The load forecasting performance of our method is evaluated with field data collected from Austin, Texas, USA.

## II. GAUSSIAN PROCESS MODEL

In this study, the response,  $y(\mathbf{x})$ , resulting from a  $d$ -dimensional input vector,  $\mathbf{x} = [x_1, \dots, x_d]^T$ , is represented as the realization of a Gaussian process [9]:

$$Y(\mathbf{x}) = \boldsymbol{\alpha}^T \mathbf{f}(\mathbf{x}) + \sigma^2 Z(\mathbf{x}) \quad (1)$$

where  $\boldsymbol{\alpha}^T \mathbf{f}(\mathbf{x})$  is the mean value of the Gaussian process (wherein  $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \dots, f_d(\mathbf{x})]$  is the basis function vector and  $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_d]$  is vector of coefficients to be estimated),  $\sigma^2$  is the process variance and  $Z(\mathbf{x})$  is a zero-mean, unit variance stationary Gaussian process. The Gaussian process,  $Z(\mathbf{x})$ , is described using an auto-correlation function,  $R = R(|\mathbf{x} - \mathbf{x}'|; \boldsymbol{\theta})$ , and associated hyperparameters,  $\boldsymbol{\theta}$ , such that:

$$E[Z(\mathbf{x})Z(\mathbf{x}')] = R(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = R(|\mathbf{x} - \mathbf{x}'|; \boldsymbol{\theta}) \quad (2)$$

A common choice for the basis functions in Gaussian process models is the use of polynomials of low order. A linear model is used in this study. There are also several common covariance functions or their combinations that can be used.

Let us defined a set of  $n$  observations (defining the training set) that form the input matrix,  $\mathbf{X} = [\mathbf{x}^1, \dots, \mathbf{x}^n]$ , and the corresponding output vector,  $\mathbf{Y} = [y(\mathbf{x}^1), \dots, y(\mathbf{x}^n)]^T$ . From this, we form the basis matrix,  $\mathbf{F} = [\mathbf{f}(\mathbf{x}^1), \dots, \mathbf{f}(\mathbf{x}^n)]^T$ , and correlation matrix  $\mathbf{R}$ , with  $ij$ -element defined as  $R^{ij} = R(\mathbf{x}^i, \mathbf{x}^j)$ ,  $i, j = 1, \dots, n$ . For any new input vector,  $\mathbf{x}$ , let  $\mathbf{r}(\mathbf{x}) = [R(\mathbf{x}, \mathbf{x}^1), \dots, R(\mathbf{x}, \mathbf{x}^n)]^T$ , represent the vector defining correlation between the new input and each of the elements of  $\mathbf{X}$ .

Note that  $Y$  follows a normal distribution such that the likelihood function may be expressed as:

$$L(\mathbf{Y}|\boldsymbol{\alpha}, \sigma, \boldsymbol{\theta}) = \frac{\det \mathbf{R}^{-0.5}}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{F}\boldsymbol{\alpha})^T \mathbf{R}^{-1}(\mathbf{Y}-\mathbf{F}\boldsymbol{\alpha})\right] \quad (3)$$

The first step is to estimate the correlation function parameters using:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \min (-\log L(\mathbf{Y}|\boldsymbol{\alpha}, \sigma, \boldsymbol{\theta})) \\ &= \arg \min \left( \frac{1}{2} \log(\det \mathbf{R}) + \frac{n}{2} \log(2\pi\sigma^2) + \frac{n}{2} \right) \end{aligned} \quad (4)$$

Given the correlation function, the generalized least-squares estimates of  $\boldsymbol{\alpha}$  and  $\sigma$  are given by:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\boldsymbol{\theta}) = (\mathbf{F}^T \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{R}^{-1} \mathbf{Y} \quad (5)$$

$$\hat{\sigma}^2 = \sigma^2(\boldsymbol{\theta}) = \frac{1}{n} (\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\alpha}})^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\alpha}}) \quad (6)$$

The mean and variance of the Gaussian process predictors are then expressed as:

$$\mu_{\hat{Y}}(\mathbf{x}) = \hat{\boldsymbol{\alpha}}^T \mathbf{f}(\mathbf{x}) + \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\alpha}}) \quad (7)$$

$$\sigma_{\hat{Y}}^2(\mathbf{x}) = \hat{\sigma}^2 [1 + \mathbf{u}^T (\mathbf{F}^T \mathbf{R}^{-1} \mathbf{F})^{-1} \mathbf{u} - \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x})] \quad (8)$$

where  $\mathbf{u} = \mathbf{F}^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}) - \mathbf{f}(\mathbf{x})$

The accuracy of the Gaussian process model depends on the selected covariance function. While there is no systematic way to choose the most appropriate covariance function, pre-analyzing trends in the data can help to select appropriate kernel functions. It is also noted that there are many common Kernel functions available [9] and that can be used in combination and offer considerable flexibility in defining predictive Gaussian process models.

## III. DATA

We collected data on maximum daily electricity loads (demand) and on maximum daily outdoor temperature for Austin, Texas, USA from 2002 to 2019. Other factors such as the local population density and economic conditions, for instance, can also influence the electricity consumption; however, these are not considered in the present study.

### A. Temperature Data

Time series of the daily maximum temperature at 2 m above ground level (AGL) are acquired from Automatic Sensing Observing Systems (ASOS) stations in Austin. The Austin ASOS site KATT (lat. 30.32°, lon. -97.76°) was selected because it is in a region of higher urban density compared to other ASOS sites in the metro area.

### B. Electricity Demand Data

Electricity demand data was provided by the Electric Reliability Council of Texas [18] (ERCOT). These data are from the period from 2002 to 2019. Available hourly data are subdivided by 8 weather zones: North (N), North-Central (NC), Far West (FW), West (W), East (E), South-Central (SC), Coast (C) and Southern (S). Because this study seeks to develop load forecasting models for Austin alone, data from the SC zone (that covers the metro areas of Austin as well as San Antonio, not of interest here) are used.

Figure 1 shows daily peak electricity demand versus daily peak temperature over three months (June, July and August) in Austin over the period from 2002 to 2019. We note that temperature and demand are strongly correlated (with a Pearson's correlation coefficient of around 0.8). This suggests that a linear model for daily peak demand prediction given temperature values might suffice; however, it is clear that there is considerable uncertainty in the data. The variability in the energy demand at a specified value of temperature (e.g., 35°C) can be quite large. Accordingly, a nonlinear model may be preferred or to improve predictions of energy demand, one might consider including other factors such as population demographics, economic state, etc. While these different factors are not considered, with appropriate choices for selected kernel functions, we employ a Gaussian process model to attempt to establish local variation in energy demand due to changes in temperature using ASOS and ERCOT data sets.

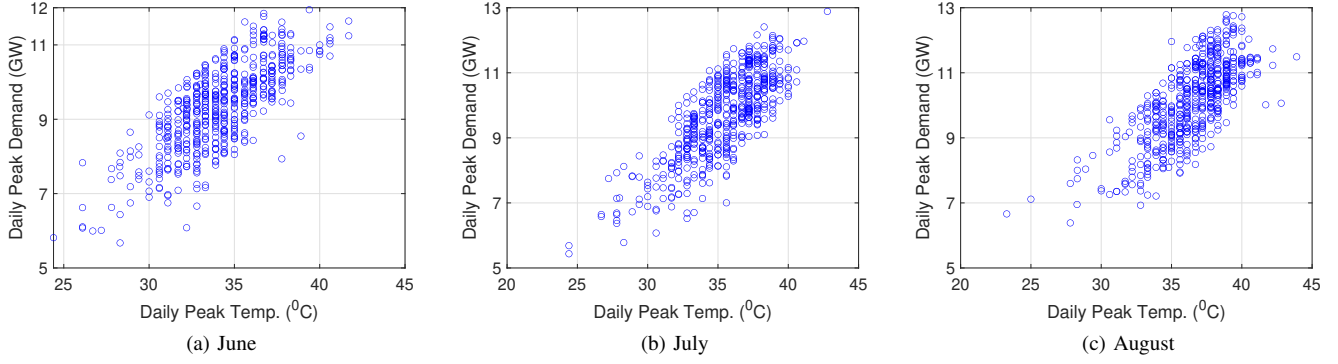


Fig. 1: Daily Energy Demand versus Daily Peak Temperature for Summer Months (June, July and August), 2002 to 2019.

#### IV. LOAD PREDICTION

A Gaussian process model is applied to predict electricity demand in two summer months (July and August) based on observed values of daily peak temperature (i.e., temperature is considered as the only external input for predicting electricity demand). In Eq. 1,  $Y$  is the electricity demand and  $\mathbf{x}$  is temperature. Note that for future forecasting where temperature is to be considered unknown at the time of prediction, we must rely on an estimate of future temperature.

In this study, we present model predictions for two cases: (i) when the actual temperature is known; and (ii) when this input variable (temperature) is not available in a future planning scenario. In both cases, we assume that historical data on peak temperature and peak electricity demand are available as part of the training set in our model prediction.

##### A. Model Based on Available Temperature Data

The data set of daily maximum temperature and load from 2002 to 2017 are divided into training and testing periods. The training period is used to establish the Gaussian process model while the testing period is used for model validation. This optimal model is further evaluated using the data from 2018 and 2019.

As kernel function, based on preliminary analysis of the data pattern, a combination of squared exponential, periodic and linear functions is used in this study:

$$R(x, x', \theta) = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right) + \theta_3 \exp\left(-\frac{\sin^2 \pi \left(\frac{x - x'}{\theta_4}\right)}{2\theta_5^2}\right) + \theta_6^2 x x' \quad (9)$$

As shown in Fig. 2, the data appears to exhibit some periodicity with some fluctuations; hence, the combination of a squared exponential and periodic covariance kernel functions is justified [9], [10]. Moreover, both temperature and load seem to share similar patterns; hence, a squared exponential function can be used to describe the smooth variation of electricity

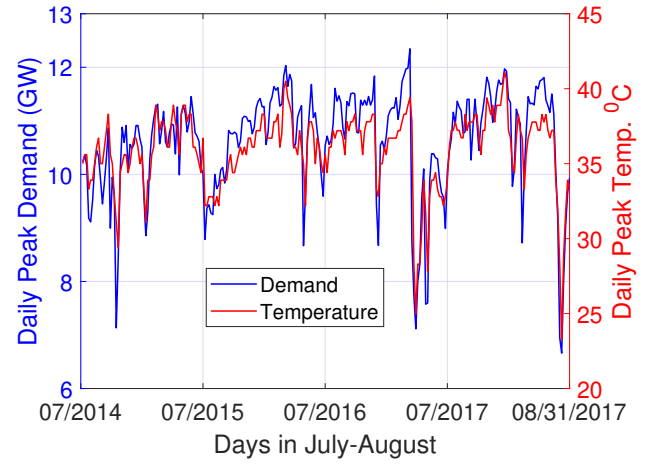


Fig. 2: Daily Observed Peak Temperature and Daily Peak Energy Demand in July and August, 2014 to 2017.

demand in term of temperature [10]. Finally, a linear kernel function component in Eq. 9 is used to model linearly varying trends in the data. As noted, the selected kernel function, then, includes a total of 6 hyper-parameters,  $\theta = [\theta_1, \dots, \theta_6]^T$ , that need to be estimated using Eq. 4.

One might expect that excessively long periods of training data will result in greater uncertainty in model prediction for any specified period. In this study, we first evaluate the prediction performance of the model by changing the length of the training data set. In the literature, there are various evaluation metrics that are commonly used; among them, error-based metrics are the most popular [11]. For each selected duration of the training set, one can estimate the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), the Root Mean Squared Error (RMSE), etc. [11], [16] for a testing set; these can all be considered as appropriate metrics for establishing the optimal size/length of the training data set. Based on preliminary analysis, patterns of MAPE, MAE and RMSE are similar for the testing sets studied; accordingly,

results only for MAPE are presented. The period of training data is considered backward in time from the testing year (e.g., if 2016 is the testing year, a 1-year training data set means the training set only includes data from the year 2015; a 2-year training data set similarly means data from 2014 and 2015 are used for training the predictive model). Here, we consider three different testing sets; they include 2016 alone, 2017 alone, and a combination of 2016 and 2017 together. For each testing case, we also vary the training period data set from 1 year to 15 years (backward in time from the testing period) and use these trained models for forecasting.

Figure 3 compares the prediction MAPE of daily peak energy demand in July-August considering different lengths of training data sets and for different testing sets. We see from Fig. 3 that including data from years prior to 2014 (i.e., lengths of training data sets greater than 3 years and 2 years with testing set at 2017 and 2016 (or 2016-2017), respectively), leads to an increase in MAPE of model predictions. As such, one might conclude that data from years prior to 2014 should not be included in the predictive Gaussian process model; two or three of the most recent years of data are sufficient for training. For example, prediction of the peak energy demand in 2017 achieves the lowest MAPE (around 3.2%) if data from only the prior three years (i.e., 2014-2016) are used to train the forecasting model. By introducing additional (older) training data, the error-based metric (MAPE) increases almost linearly. We should note that there are other factors that are not accounted in the model; compared to the changes in temperature, such factors could have a significant influence on the peak energy demand before and after 2014. Also, with the choice of kernel function in Eq. 9, there are only 6 unknown parameters to be estimated; hence, three years of data (with 62 data points for each year for the months of July and August) are sufficient.

Figure 4 shows model prediction results for July and August of 2017 where three years of data from 2014 to 2016 are selected as the training set. Actual and predicted time series of daily peak demand are presented in Fig. 4a; it can be seen that the mean prediction matches well with observed values (as discussed before, the MAPE value is around 3%). Also, the 95% confidence interval (CI) on predicted peak energy demand (see Fig. 4b) adequately covers all observed peak energy demand levels and the Prediction Interval Coverage Probability [19] is virtually 100%. As was seen in Fig. 1, at any specified peak daily temperature, the bandwidth on historical peak energy demand is greater than 2 GW; in contrast, Fig. 4b suggests that the width of the 95% CI on predicted peak energy demand based on the Gaussian process model is around 1.88 GW. Interestingly, model predictions for July are somewhat better than for August. Additional model evaluations for 2018 and 2019 are presented in Figs. 5 and 6, respectively. Similar conclusions to the testing for 2017 can be drawn for 2018 and 2019 as one can see in these figures. By testing the proposed model over these three different periods, one can conclude that the Gaussian process model appears to be robust, easy to implement, and accurate for estimating the

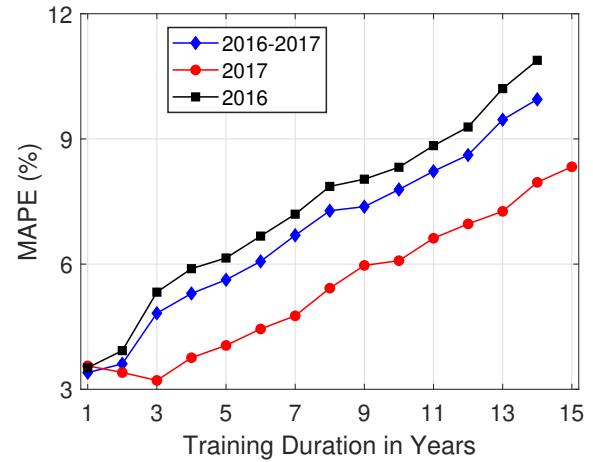


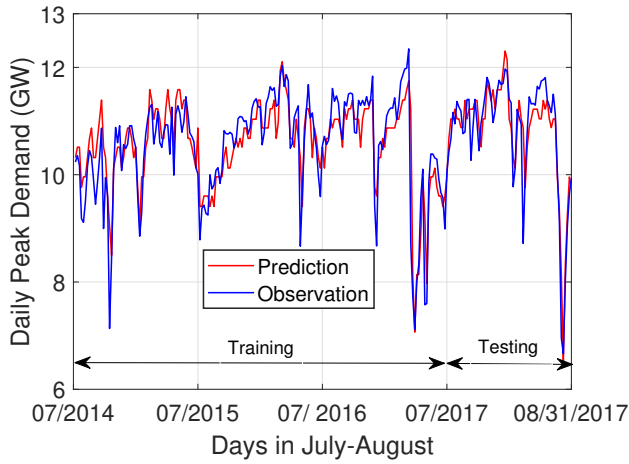
Fig. 3: Prediction MAPE of Peak Energy Demand in July-August for Different Testing Sets (blue curve: two testing for 2016-2017; red curve: testing for 2017; black curve: testing for 2016) and with Varying Amounts of Training Data (Training Duration in years is Accounted for Backward from the Testing Period).

daily peak electricity demand given temperature.

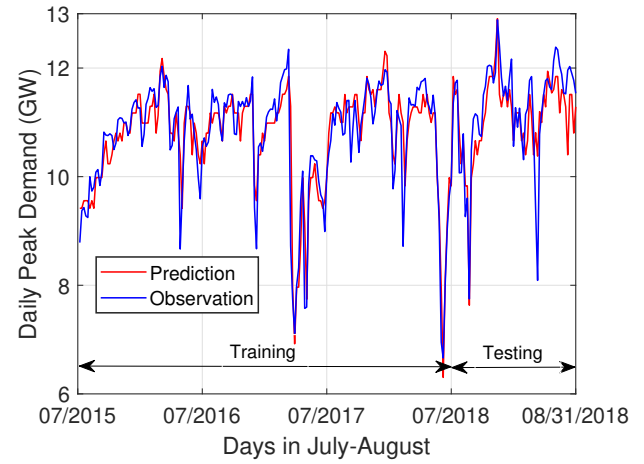
#### B. Scenario Analysis for Peak Energy Demand Prediction when Future Temperature is Unknown

Since, in general, for future peak energy demand prediction, temperature data may not be available, a predictive model for temperature must first be established before peak energy demand can be predicted. A data-driven model such as a Gaussian process model or a physical climate model may be used to predict future temperatures. Then, those predicted temperatures may be introduced into the peak energy demand forecasting model to predict the future energy consumption. Note that the projected temperature values will be associated with considerable uncertainty [19]; the challenge is how to propagate this uncertainty into that on predicted electricity demand. We acknowledge that this task is challenging; accordingly, instead of developing either any statistical or physical model for temperature forecasting, we employ a “scenario analysis” framework to assess only probabilistic characteristics and uncertainty in electricity demand prediction.

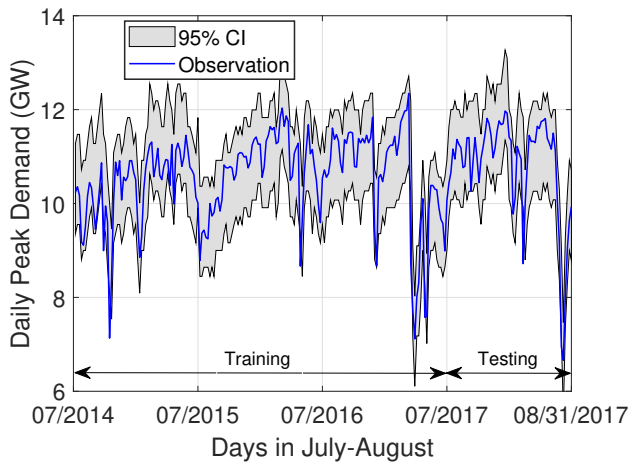
Assume that temperature in a specified year of interest—2020, for example—is similar to any single past year in the period from 2002 to 2019. Temperature of each year in the historical data set is considered as the “actual” temperature in year to be predicted (2020); this temperature is then used as input for load forecasting in 2020 using the Gaussian process model. This process is repeatedly applied considering each of the historical years separately. Given  $N$  years of data,  $N$  different scenarios of peak energy demand prediction may be thus obtained. In this manner, an uncertainty assessment in future load prediction can be carried out. The procedure of such a scenario-based analysis can be summarized as follows:



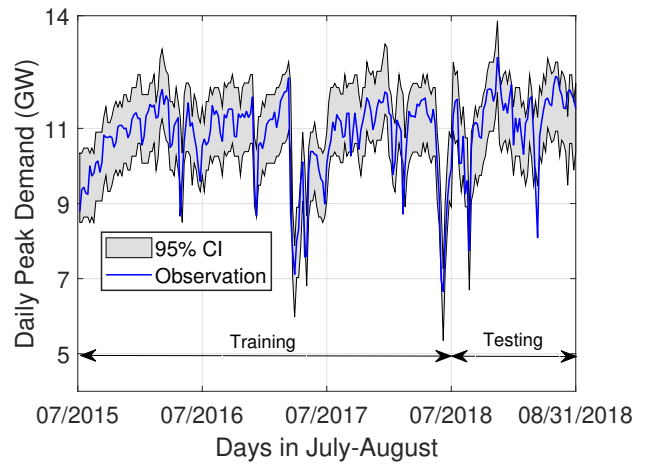
(a) Time Series Mean Prediction



(a) Time Series Mean Prediction



(b) 95% CI



(b) 95% CI

Fig. 4: Electricity Demand Prediction in July and August for 2017. Data of July and August in the three most recent years (2014-2016) are used for training.

Fig. 5: Electricity Demand Prediction in July and August for 2018. Data of July and August in the three most recent years (2015-2017) are used for training.

- Step 1: Assume the temperature time series in a future year of interest is similar to that for a single year in the historical sample.
- Step 2: Use the Gaussian process model developed to predict future peak energy demand in that future year (for the 62 days in the July-August only).
- Step 3: The assumption that the temperature time series for the future year is the same as that of any prior year in the historical sample may not be reasonable or realistic. Accordingly, we do not make predictions for point-estimate peak energy demand at the specific time in the future; instead, we seek multiple estimates of future energy demand and attempt to quantify the uncertainty in this demand. After estimating the demand for the 62 days (in July and August), we sort the peak energy demand predictions from largest to smallest and rank them from 1

to 62. This is equivalent to assessing prediction quantiles of energy demand.

- Step 4: We repeat such predictive energy demand quantile computations by considering all single-year temperature time series in the historical sample. In this manner, we can quantify the uncertainty in peak energy demand for a future year without any temperature data; this uncertainty accounts for both variability in July-August temperature scenarios as well in the Gaussian process model prediction.

The results from such a scenario-based analysis for 2017, 2018 and 2019 are presented in Fig. 7. In applying the Gaussian process model, three years of data collected from 2014 to 2016 are used to train the model. In Figs. 7a, 7b and 7c, the light blue thin curves represent projected peak loads in 2017, 2018 and 2019, respectively, where each curve shows



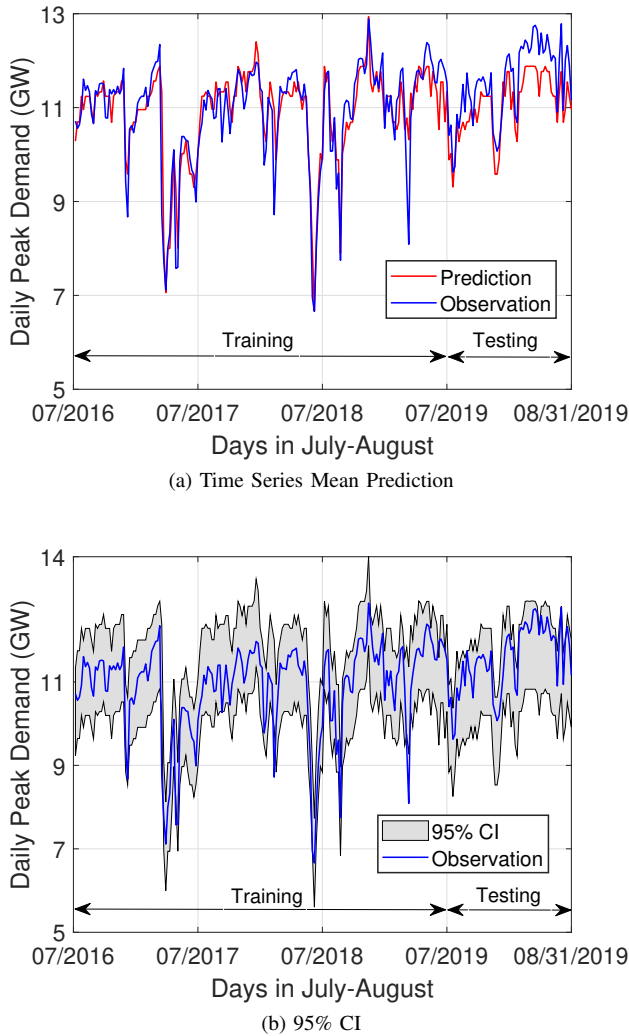


Fig. 6: Electricity Demand Prediction in July and August for 2019. Data of July and August in the three most recent years (2016-2018) are used for training.

sorted energy demand values over 62 days and corresponds to a distinct scenario where a past year's temperature time series (starting from 2002) is assumed to be applied for the projected future year of interest. The red curve is obtained when the actual temperature of the projected future year is used, while the dark blue curve represents the actual observed peak energy demand for the year of interest (note that this last curve serves to validate the Gaussian process model as well as to assess the scenario-based analysis). It can be confirmed that the scenario-based peak energy demand predictions are comparable to the peak energy demand values actually experienced, but the predicted bandwidth on peak energy demand (if one considers all scenarios as well as all of the 62 ranks in July and August of the projected year of interest) is significantly wider compared to the case when actual temperature for the year on interest is known. Thus, the ability to project accurate temperatures

at the time of interest plays a vital role in energy demand forecasting.

A comparison of all the maximum (of 62) peak energy demand values (Rank 1) are presented in Figs. 7d, 7e and 7f. With Fig. 7d, for instance, we see that the lowest and highest scenario-based results are 11.10 GW and 13.59 GW, respectively, compared with the actual value of 11.97 GW. The 2017-scenario prediction is seen to be the best while the 2019 prediction is associated with the largest error. This may be partially explained when we study Fig. 8, which compares actual energy demand data patterns over the three different years of interest (2017, 2018 or 2019) with demand projections based on scenarios using the historical data. First, it can be seen that the distribution of daily peak temperatures in 2017, 2018 and 2019 are quite similar, ranging from around 32°C to 40°C; these are somewhat higher than temperatures in the historical sample that ranged from 27°C to 42°C. This explains why uncertainty in peak daily energy demand in July and August at the lower end (rank 62) of the scenario-based predictions is greater than it is at the upper end (see Figs. 7a, 7b and 7c). A second point worth noting is that, from 2017 to 2019, although temperatures in Austin are quite similarly distributed, demand tends to steadily increase with time over that period. Note that there are other factors related to economic activity, household usage, etc. that can significantly influence electricity consumption but these are not accounted for in the model presented in this study. The larger error in peak energy demand prediction in 2019 compared to 2017 may be partly explained by such unaccounted for influences.

## V. CONCLUSIONS

In this study, a Gaussian process model was developed to forecast daily peak electricity demand in Austin, Texas, USA. Based on a preliminary investigation of data patterns (daily load versus temperature), an appropriate covariance kernel function was proposed to represent characteristics of data. Error-based metrics were used to select the optimal duration (length) of the training data set. The performance of the proposed model was evaluated in two cases—whether future temperature in the year of interest is available or not. When temperature data at the time of prediction is available, the Gaussian process model is seen to offer an easy and robust approach for forecasting energy demand at the city scale. Over different testing data sets, the error metric (MAPE) used to assess prediction accuracy was relatively small (around 3%), whereas another metric (PICP) that considers the 95% CI was 100% with a reasonable bandwidth (around 1.88 GW) on the peak energy demand. A scenario-based framework was also illustrated for assessing the uncertainty in peak energy demand prediction in the case when future actual temperature data are unknown or unavailable. Results showed that uncertainty in this case is significantly larger than when temperature data are available, demonstrating the strong influence of accurate temperature as a predictor of energy demand. In addition, other factors can strongly influence trends noted in increasing

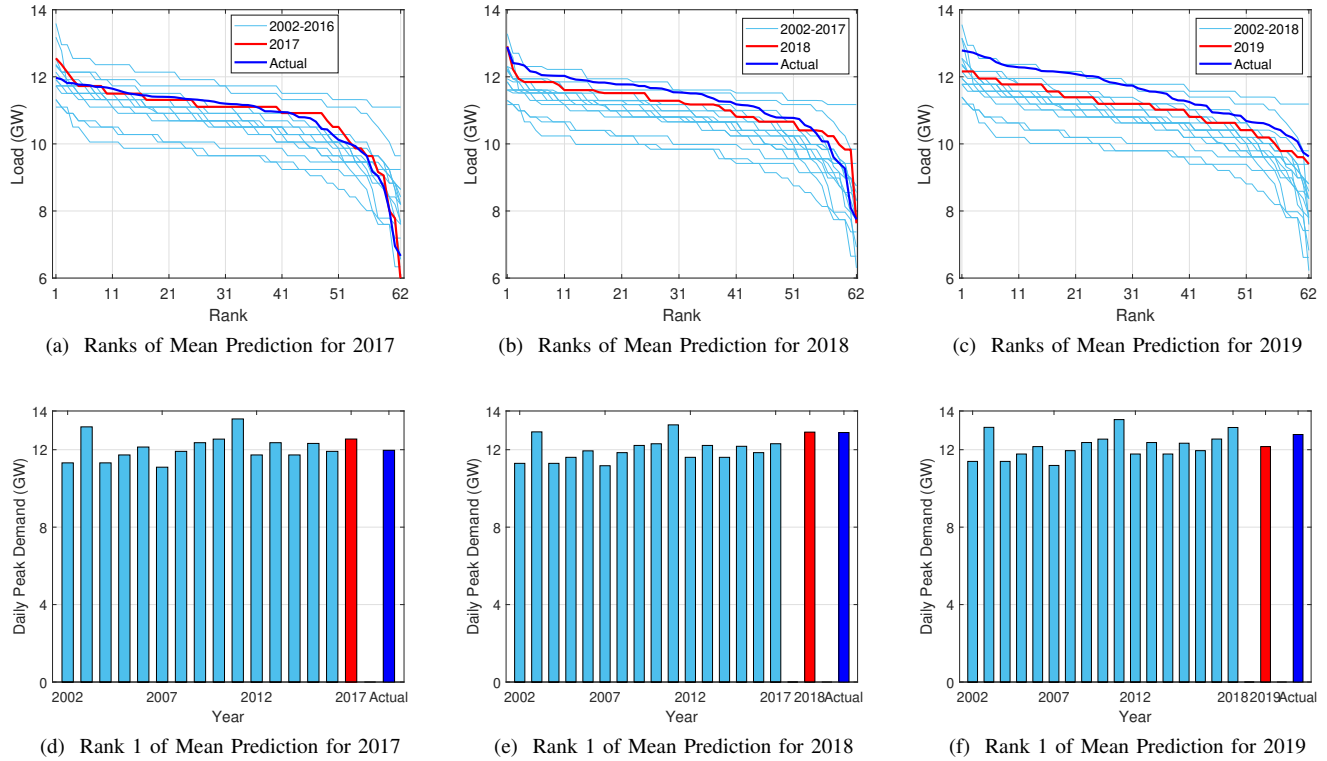


Fig. 7: Scenario-Based Peak Energy Demand Prediction for 2017 (first column), 2018 (second column) and 2019 (third column). Top row: Each Curve shows 62 Point Predictions for the Days in July and August that are Sorted in Descending Order, Bottom row: Comparison of the Highest Values Corresponding to each Curve in the Top Figures. Light blue curves/bars: temperature time series for each year in the historical sample are used as input for the predictive model; red curves/bars: when the actual temperature of the predicted year (first column: 2017, second column: 2018, last column: 2019) is used for demand prediction; dark blue curve/bar: observed peak energy demand for the predicted year (for validation).

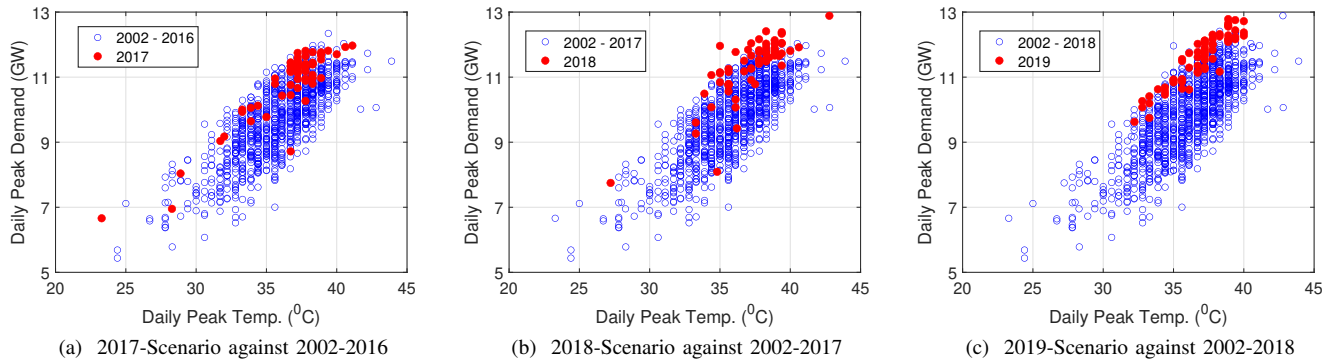


Fig. 8: Daily Observed Peak Energy Demand versus Daily Peak Temperature in different Scenarios: Comparison of data patterns in the year of interest (a: 2017, b: 2018 and c: 2019) versus projections using scenarios with historical temperature data. The blue dots represent scenarios using the historical data from 2002 to the year of interest (a: 2017, b: 2018 and c: 2019), while the red dots represent observed data for the year of interest (a: 2017, b: 2018 and c: 2019).

electricity consumption in recent years; these other factors can introduce large uncertainty in proposed models such as the one developed here, which only explicitly considered temperature

as an input in the Gaussian process. Future topics for investigation could consider incorporating temperature forecasting models into energy demand predictive models and could also

consider taking into account other factors beyond temperature that can play a role in energy demand prediction.

#### ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. CMMI-1663044. The authors are grateful for this support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

#### REFERENCES

- [1] D. W. Bunn, "Forecasting loads and prices in competitive power markets," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 163–169, 2000.
- [2] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.
- [3] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *IEEE Transactions on Power Systems*, vol. 16, no. 3, pp. 498–505, 2001.
- [4] S. S. Pappas, L. Ekonomou, D. C. Karamousantas, G. Chatzarakis, S. Katsikas, and P. Liatsis, "Electricity demand loads modeling using autoregressive moving average (ARMA) models," *Energy*, vol. 33, no. 9, pp. 1353–1360, 2008.
- [5] S. S. Pappas, L. Ekonomou, P. Karampelas, D. Karamousantas, S. Katsikas, G. Chatzarakis, and P. Skafidas, "Electricity demand load forecasting of the Hellenic power system using an ARMA model," *Electric Power Systems Research*, vol. 80, no. 3, pp. 256–264, 2010.
- [6] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse, "Methodology for long-term prediction of time series," *Neurocomputing*, vol. 70, no. 16–18, pp. 2861–2869, 2007.
- [7] G. Dudek, "Pattern-based local linear regression models for short-term load forecasting," *Electric Power Systems Research*, vol. 130, pp. 139–147, 2016.
- [8] H.-Z. Li, S. Guo, C.-J. Li, and J.-Q. Sun, "A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm," *Knowledge-Based Systems*, vol. 37, pp. 378–387, 2013.
- [9] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.
- [10] A. K. Prakash, S. Xu, R. Rajagopal, and H. Y. Noh, "Robust building energy load forecasting using physically-based kernel models," *Energies*, vol. 11, no. 4, p. 862, 2018.
- [11] Y. Yang, S. Li, W. Li, and M. Qu, "Power load probability density forecasting using Gaussian process quantile regression," *Applied Energy*, vol. 213, pp. 499–509, 2018.
- [12] A. T. Eseye, M. Lehtonen, T. Tukia, S. Uimonen, and R. J. Millar, "Machine learning based integrated feature selection approach for improved electricity demand forecasting in decentralized energy systems," *IEEE Access*, vol. 7, pp. 91 463–91 475, 2019.
- [13] G. Zhang and J. Guo, "A novel method for hourly electricity demand forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1351–1363, 2019.
- [14] M. S. Al-Musaylh, R. C. Deo, J. F. Adamowski, and Y. Li, "Short-term electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF reanalysis atmospheric predictors in southeast Queensland, Australia," *Renewable and Sustainable Energy Reviews*, vol. 113, p. 109293, 2019.
- [15] P. Pelka and G. Dudek, "Pattern-based forecasting monthly electricity demand using multilayer perceptron," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2019, pp. 663–672.
- [16] P. Jiang, R. Li, N. Liu, and Y. Gao, "A novel composite electricity demand forecasting framework by data processing and optimized support vector machine," *Applied Energy*, vol. 260, p. 114243, 2020.
- [17] M. Hekkenberg, R. Benders, H. Moll, and A. S. Uiterkamp, "Indications for a changing electricity demand pattern: The temperature dependence of electricity demand in the Netherlands," *Energy Policy*, vol. 37, no. 4, pp. 1542–1551, 2009.
- [18] ERCOT. (2019) Ercot. <http://www.ercot.com>. [Online; accessed 16-December-2019].
- [19] D. E. Jahn, W. A. Gallus, P. T. Nguyen, Q. Pan, K. Cetin, E. Byon, L. Manuel, Y. Zhou, and E. Jahani, "Projecting the most likely annual urban heat extremes in the Central United States," *Atmosphere*, vol. 10, no. 12, p. 727, 2019.

# An Experimental On Heat Transfer Characteristics Of The Cascade Heat Exchanger In Refrigeration System Using R32/CO<sub>2</sub>

Thanhtrung Dang

Department of Thermal Engineering  
HCMC University of Technology and  
Education (HCMUTE)

Ho Chi Minh City, Vietnam  
trungdang@hcmute.edu.vn

Vanloi Nguyen

Department of Thermal Engineering  
HCMC University of Technology and  
Education (HCMUTE)

Ho Chi Minh City, Vietnam  
nguyenloi.cdnqn@gmail.com

Hoangtuan Nguyen

Faculty of Refrigeration  
College of Technology II

Ho Chi Minh city, Vietnam  
tuannghuyenhoang@hvct.edu.vn

**Abstract**—An experimental on heat transfer characteristics of the cascade heat exchanger in a refrigeration system was investigated. In this paper, the cascade system used the couple of R32/CO<sub>2</sub> as refrigerant: CO<sub>2</sub> in low stage cycle, R32 in high stage cycle. The heat exchanger is double-tube type with arranging the spiral shape, with the inner diameters are 4 mm and 8 mm for the inner tube and the outer tube, respectively. Using a microchannel evaporator, this system was operated with the evaporative temperature of -26°C. In this study, the temperature profile and the thermal capacity were determined. Moreover, the results obtained from the experimental data are good agreement with the theoretical calculations. These results are very important to calculate and design the cascade heat exchanger using CO<sub>2</sub>.

**Keywords**—cascade, refrigeration, heat exchanger, R32, CO<sub>2</sub>

## I. INTRODUCTION

Nowadays, refrigeration technology is deeply applied in many fields such as industry, agriculture, healthcare,... For cooling at low temperature, people usually use multi-stage refrigeration system. Most of refrigeration systems used Hydrochlorofluorocarbon (HCFC) or Hydrofluorocarbon (HFC) as refrigerants. However, when the refrigerants leak to the atmosphere, they cause the ozone depletion potential and global warming potential. Carbon dioxide is one of the refrigerants of the future, easy to produce, environmentally friendly as well as good thermodynamic properties for low temperature refrigeration.

In the researches about transcritical CO<sub>2</sub> refrigeration system, Zhang et al. [1] analyzed theoretically the effect of the internal heat exchanger (IHE) on the performance of the ejector expansion transcritical CO<sub>2</sub> refrigeration system, based on the first law of thermodynamics. This study found that the addition of IHE in the CO<sub>2</sub> ejector refrigeration cycle increases the ejector entrainment ratio and the ejector efficiency, and decreases pressure recovery under the same gas cooler pressures. Gupta et al. [2] carried out a simulation based study to analyze the performance of modified CO<sub>2</sub> transcritical refrigeration system with work recovery turbine. This paper emphasizes upon design and operating parameters based on local environmental conditions for the best possible performance. In addition, Xu et al. [3] experimented results on a performance comparison of a transcritical CO<sub>2</sub> ejector system without an internal heat exchanger and a transcritical CO<sub>2</sub> ejector system with an internal heat exchanger. In this study, the experimental results are used to validate the findings that the internal heat exchanger weakens the

contribution of the ejector to the system performance. Song et al. [4] researched the adaptation of the capillary tube in the transcritical CO<sub>2</sub> refrigeration system by a separated flow model of a CO<sub>2</sub> capillary. The capillary based transcritical CO<sub>2</sub> system could achieve performance close to that of electronic expansion valve based system in a wide range of gas-cooler outlet temperature. Besides, Yang et al. [5] reviewed optimal high pressure correlations of transcritical CO<sub>2</sub> cycles and found a potential significant COP loss near the critical pressure due to error propagation from optimal high pressure approximation to actual COP. Tao et al. [6] investigated system for the transcritical CO<sub>2</sub> residential air conditioning with an internal heat exchanger on the system coefficient of performance (COP) for working conditions. The COP reduces about 20% when gas cooler side air inlet temperature increases from 32.5°C to 37°C, augments about 27% when the air inlet velocity of the gas cooler increases from 0.68 m/s to 1.8 m/s and augments about 11% when evaporation temperature increases from 8.7°C to 13.9°C. In the researches about subcritical CO<sub>2</sub> refrigeration system, Lee et al. [7] analyzed a cascade refrigeration system that uses carbon dioxide and ammonia as refrigerants, to determine the optimal condensing temperature of the cascade condenser given various design parameters, to maximize the COP and to minimize the exergy destruction of the system. Ma et al. [8] presented a CO<sub>2</sub>/NH<sub>3</sub> cascade refrigeration system, in which a falling film evaporator–condenser is used as the cascade heat exchanger. Mohammadi and Ameri [9] conducted a comparative study of six configurations of an absorption–compression cascade refrigeration system. At specific discharge pressure, a system with inter-cooler, after-cooler and double-effect absorption chiller would result in the highest improvement of the system. Cai et al. [10] proposed a comparison based on the analysis of the properties of carbon dioxide. A model of open carbon dioxide refrigeration system is developed, and the relationship between the storage environment of carbon dioxide and refrigeration capacity is conducted. In this study, the refrigeration capacity loss by heat transfer in supercritical state is much more than that in two-phase region and the refrigeration capacity loss by remaining carbon dioxide has little relation to the state of CO<sub>2</sub>. Li et al. [11] presented an experimental data of CO<sub>2</sub> flow condensation heat transfer for mass fluxes ranging from 100 to 500 kg/m<sup>2</sup>s inside a 4.73 mm inside diameter, smooth horizontal copper tube, at saturation temperatures between -10°C and 0°C under a wide range of vapor quality conditions. The results showed that when the test mass flux was greater than or equal to 300 kg/m<sup>2</sup>s, for vapor qualities greater than 0.4. The heat transfer rate increases with increasing mass flux and vapor quality, and

also increases with decreasing saturation temperature. In addition, Patel et al. [12] investigated a cascade refrigeration system operating with CO<sub>2</sub> in the low temperature cycle and NH<sub>3</sub> as well as C<sub>3</sub>H<sub>8</sub> in the high temperature cycle for the thermo-economic optimization. Optimization results are used for the comparative analysis of both the refrigerant pairs (NH<sub>3</sub>/CO<sub>2</sub> and C<sub>3</sub>H<sub>8</sub>/CO<sub>2</sub>). Most of researches focus on transcritical CO<sub>2</sub> refrigeration system with coefficient of performance (COP) comparisons under different conditions while the studies on subcritical CO<sub>2</sub> refrigeration system are limited. In fact, to achieve high COP we must operate the system under critical point, cascade system is good solution for this case. NH<sub>3</sub>, C<sub>3</sub>H<sub>8</sub>, R404A are refrigerants that chosen for high stage cycle meanwhile Refrigerant R32 is one of the most environmentally friendly option and less studied than the others. Thus, this study investigates heat transfer characteristics on pair of R32/CO<sub>2</sub> in cascade refrigeration system.

From literature reviews above, most of researches focus on transcritical CO<sub>2</sub> refrigeration system with coefficient of performance (COP) comparisons under different conditions while the studies on subcritical CO<sub>2</sub> refrigeration system are limited. In fact, to achieve high COP we must operate the system under critical point, cascade system is good solution for this case. NH<sub>3</sub>, C<sub>3</sub>H<sub>8</sub>, R404A are refrigerants that chosen for high stage cycle meanwhile R32 is the most environmentally friendly option and less studied than the others. Thus, the investigation on the heat transfer characteristics of the cascade heat exchanger in a refrigeration system using R32/CO<sub>2</sub> is essential. The research scope is limited to a range of 1.5 kW cooling capacity and -20°C cold room temperature. The experimental system will be installed at the Hochiminh City University of Technology and Education, Vietnam.

## II. METHODOLOGY

### A. Calculation and Design.

Based on the research scope and the calculation method of a refrigeration system, the parameters of main thermodynamic points are listed in Table 1. The calculated thermal capacity of the cascade heat exchanger is 1.8 kW. The calculated power input of CO<sub>2</sub> is 433 W, so the SANDEN compressor with the capacity of 440 W was chosen. The calculated thermal capacity of the R32 condenser is 2.2 kW. The calculated power input of R32 is 448 W, so the inverter DAIKIN compressor with the capacity 1HP was chosen in this study.

The heat transfer rate was calculated from:

$$Q_t = mc_p \Delta t \quad (1)$$

where  $Q_t$  is heat transfer rate (W)  
 $m$  is mass flow rate (kg/s)  
 $c_p$  is specific heat of fluid (kJ/kg.K)  
 $\Delta t$  is temperature difference (K)

The heat balance equation in heat exchanger express by:

$$Q_t = G_1 \cdot c_{p1} \cdot (t'_1 - t''_1) = G_2 \cdot c_{p2} \cdot (t'_2 - t''_2) \quad (2)$$

where

$G_1, G_2$  are mass flow rate of hot and cold fluid (kg/s)  
 $c_{p1}, c_{p2}$  are specific heat of hot and cold fluid (kJ/kg.°C)  
 $t'_1, t'_2$  are temperature at inlet of hot and cold fluid (°C)  
 $t''_1, t''_2$  are temperature at outlet of hot and cold fluid (°C)

The total cooling capacity of evaporator was expressed as:

$$h_t = h_s + h_l \quad (3)$$

where  $h_t$  is the cooling capacity of evaporator (W)  
 $h_s$  is the sensible heat (W)  
 $h_l$  is the latent heat (W)

TABLE 1. PARAMETERS OF THE THERMODYNAMIC POINTS OF THE CYCLES

Points	CO <sub>2</sub> cycle		R32 cycle	
	t (°C)	P (bar)	t (°C)	P (bar)
1	-21	16.3	8	9.2
2	45	45.0	63	24.8
3	10	45.0	40	24.8
4	-26	16.3	4	9.2
1'	-26	16.3	4	9.2

The double-pipe type was chosen for design the cascade heat exchanger. Based on the thermal capacity, the calculated heat transfer area is 0.16 m<sup>2</sup> with the length pipe of 13 m, with the inner diameters are 4 mm and 8 mm for the inner tube and the outer tube, respectively. For CO<sub>2</sub> cycle, a microchannel evaporator was used in this test loop.

### B. Experimental setup

The experimental system is divided into two stages: Low stage and high stage. At low stage: CO<sub>2</sub> is used as a circulating refrigerant, the CO<sub>2</sub> compressor (reciprocating type) compresses CO<sub>2</sub> vapor to high pressure, this vapor continues to enter the double-pipe condenser (cascade heat exchanger), where the superheated vapor CO<sub>2</sub> releases heat and condenses a liquid state, which continues to flow through a throttle valve, depressurizes and enters the microchannel evaporator to absorb heat from the enclosed space.

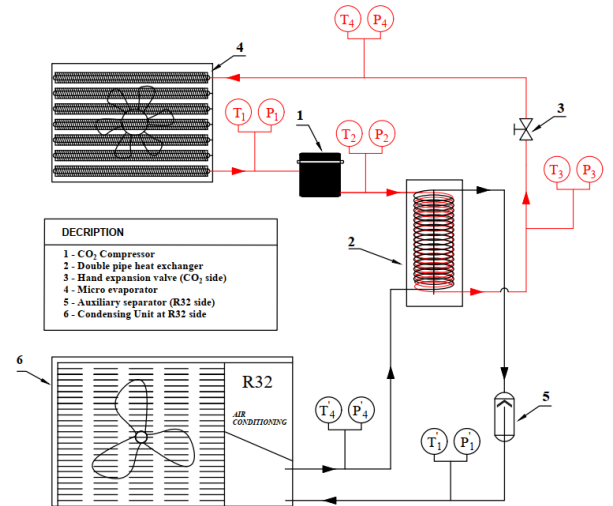


Fig. 1. Schematic of the test loop

At high stage: R32 is used as a circulating refrigerant, R32 compressor (rotary type) compresses the superheated vapor R32 to high pressure and temperature, this vapor will flow into the condenser. After condensing, R32 flows through a throttle valve to reduce pressure and enter the cascade heat exchanger



(evaporator) to receive heat released from the CO<sub>2</sub> side. This vapor continues to return the R32 compressor and the cycle goes on and on. A principle diagram is shown in Fig.1. A photo of the experimental system is shown in Fig.2 also.



Fig. 2. A photo of the test loop

Parameters such as temperature, pressure at the nodes of the system shown in Fig. 1 are collected during the operation process. The air velocity and the humidity of cooling environment (CO<sub>2</sub> side) are also collected to support for calculation. Specifically, data on surface temperatures are also collected to compare the temperature profile at different operating conditions of the system. The accuracies of the testing apparatuses are shown in Table 2.

TABLE 2. ACCURACIES AND RANGES OF TESTING APPARATUSES

Testing apparatus	Accuracy	Range
Thermometer	$\pm 0.1^{\circ}\text{C}$	$-270 - 400^{\circ}\text{C}$
Pressure sensor	$\pm 0.5\text{FS}$	$0 - 100 \text{ bar}$
Anenometer	3% FS	$0.3 - 45 \text{ m/s}$
Humidity meter	$\pm 3\% \text{ FS}$	$1.0 - 99.9\%$
Thermal IR camera	$\pm 0 - 5^{\circ}\text{C}$	$-20 - 250^{\circ}\text{C}$

### III. RESULT AND DISCUSSION

Experimental data obtained the cascade refrigeration system are under the ambient temperature from  $32$  to  $34^{\circ}\text{C}$ . This system was operated with the evaporative temperature of  $-26^{\circ}\text{C}$ . The thermodynamic points of R744 cycle on p-h diagram are shown in Fig. 3 for experimental method. From the figure, the pressure drop of the condenser is lower than that obtained from the evaporator. The difference is due to the accuracy of apparatuses, the heat exchanger type as well as the surface fouling. The results show that the experimental results are in good agreement with those obtained from the theoretical results for R744 cycle, as shown in Fig. 4. An agreement between the experimental and theoretical results is also obtained for R32 cycle. The Fig. 5 shows the thermodynamic points of R744 cycle on p-h diagram.

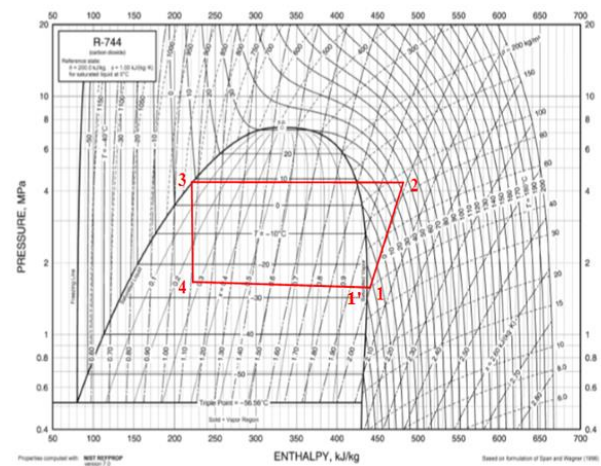


Fig. 3. The experimental points of R744 cycle on p-h diagram

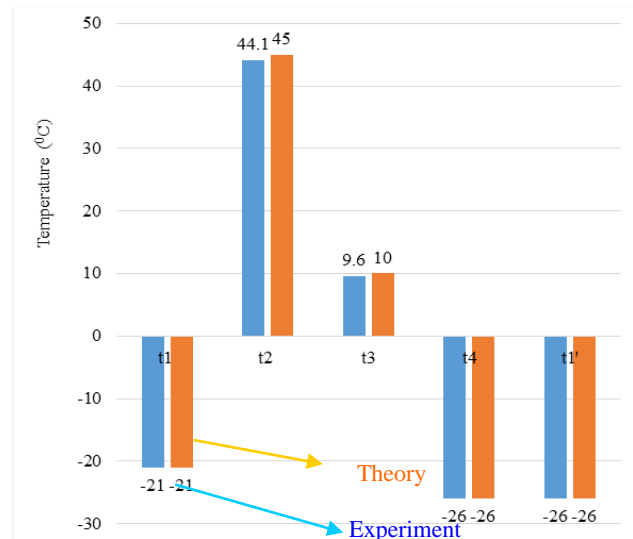


Fig. 4. Comparison between theoretical and experimental results for R744 cycle

The Fig. 6 shows the ambient temperature versus the temperature points of the cascade heat exchanger. The changing ambient temperature was changed the parameters such as the thermodynamic point, the log mean temperature difference, the heat transfer rate of the cascade heat exchanger, the compressor power input, etc.

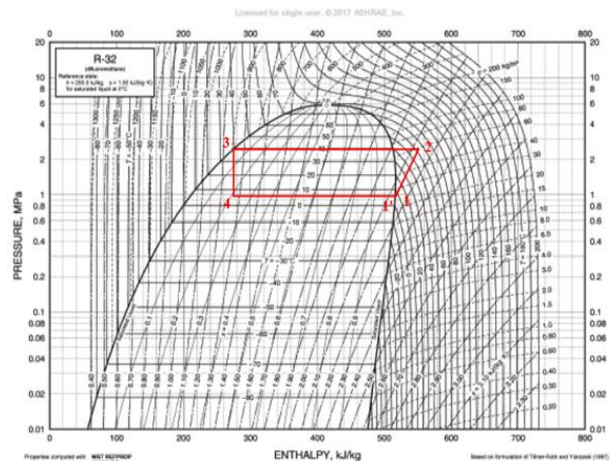


Fig. 5. The experimental points of R32 cycle on p-h diagram

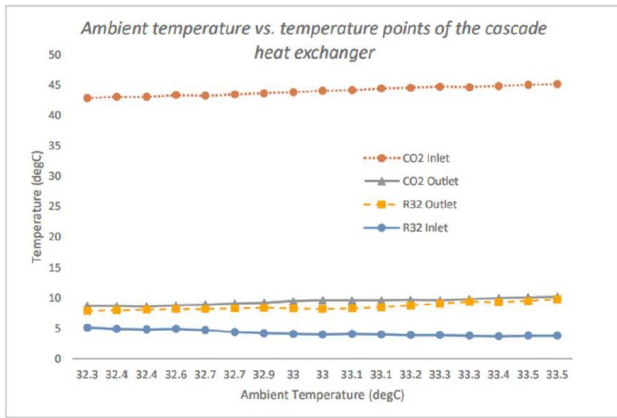


Fig. 6. Ambient temperature vs. temperature points of the cascade heat exchanger

When the ambient temperature changes from 32.3 °C to 33.5 °C, the heat transfer rate of the cascade heat exchanger also changes the value from 1546.2 W to 1899 W, as shown in Fig. 7. At an experimental temperature of 33 °C, the  $Q_{\text{cascade}}$  value of 1826.5 W (theoretically 1803 W) is consistent with the theory results.

A comparison between theory and experiment results is indicated in Table 3. The temperature distributions of the thermodynamic points of the CO<sub>2</sub> and R32 cycles are quite consistent between theory and experiment. The CO<sub>2</sub> temperature difference between the inlet and the outlet of the cascade heat exchanger by theoretical and experimental methods is not much. Typically, the inlet temperature difference is 0.9 °C (the theoretical value is 45 °C compared to the experimental value of 44.1 °C) and outlet temperature difference 0.4 °C (the theoretical value is 10 °C compared to the experimental value of 9.6 °C). The cold room temperature has reached lower than the design temperature -20°C (actually reached -25 °C).

The above results have contributed significantly to the calculation and design of the cascade refrigeration system using CO<sub>2</sub> and another refrigerant. At the same time, the study also gives the most general overview of the cascade refrigeration system using CO<sub>2</sub> at subcritical temperatures.

TABLE 3. COMPARISON BETWEEN THEORY AND EXPERIMENT RESULTS

Parameters	CO <sub>2</sub>		R32	
	Theory	Experiment	Theory	Experiment
Evaporative temperature, °C	-26	-26	4	4
Condensing temperature, °C	10	9.6	40	39.4
Inlet temperature of the cascade, °C	45	44.1	4	4
Outlet temperature of the cascade, °C	10	9.6	8	8.2
Power input of compressor, W	433	343.2	448	411.8

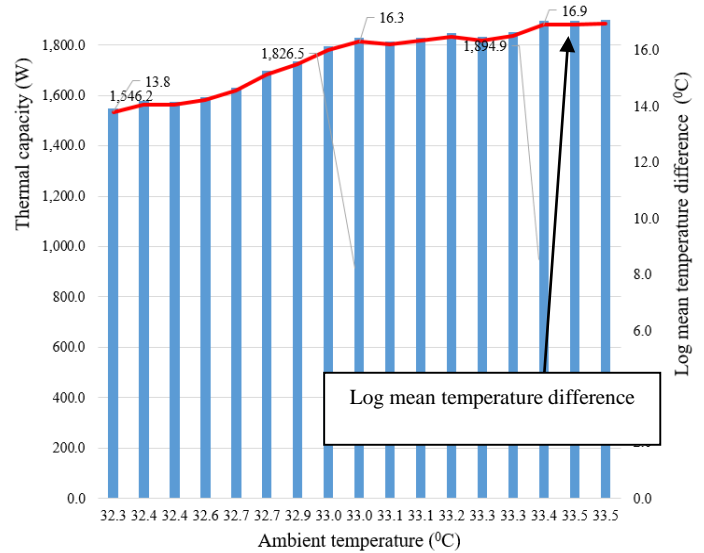


Fig. 7. Ambient temperature vs. thermal capacity of the cascade heat exchanger

#### IV. CONCLUSION

The heat transfer characteristics of the cascade heat exchanger in refrigeration system using R32/CO<sub>2</sub> have experimented in this study. The main results can be summarized as follows:

- The changing ambient temperature has changed system parameters such as the thermodynamic points, the log mean temperature difference, the heat transfer rate of cascade heat exchanger, the compressor power input, etc.
- When the ambient temperature changes from 32.3 °C to 33.5 °C, the heat transfer rate of the cascade heat exchanger also changes the value from 1546.2 W to 1899 W.
- The temperature distributions of the thermodynamic points of the CO<sub>2</sub> and R32 cycles are quite consistent between theory and experiment.
- The CO<sub>2</sub> temperature difference between the inlet and the outlet of the cascade heat exchanger by theoretical and experimental methods is not much. Typically, the inlet temperature difference is 0.9 °C (the theoretical value is 45 °C compared to the experimental value of 44.1 °C) and outlet temperature difference 0.4 °C (the theoretical value is 10 °C compared to the experimental value of 9.6 °C).
- The above results have contributed significantly to the calculation and design of the cascade refrigeration system using CO<sub>2</sub> and another refrigerant.

#### ACKNOWLEDGEMENTS

The supports of this work by the project No. B2020-SPK-04 (sponsored by the Vietnam Ministry of Education and Training) are deeply appreciated.

# REFERENCES

- [1] Zhen-ying Zhang, Yi-tai Ma, Hong-li Wang, Min-xia Li, Theoretical evaluation on effect of internal heat exchanger in ejector expansion transcritical CO<sub>2</sub> refrigeration cycle. *Applied Thermal Engineering* 50, 932 – 938 (2013).
- [2] Dileep Kumar Gupta, Mani Shankar Dasgupta, Performance of CO<sub>2</sub> Trans-Critical Refrigeration System with Work Recovery Turbine in Indian Context. *Energy Procedia* 109, 102 – 112 (2017).
- [3] Xiao-xiao XU , Guang-ming CHEN, Li-ming TANG, Experimental evaluation of the effect of an internal heat exchanger on a transcritical CO<sub>2</sub> ejector system. *Applied Physics & Engineering* 12, 146 – 153 (2011).
- [4] Yulong Song, Jing Wang, Feng Cao, Investigation on the adaptivity of the Transcritical CO<sub>2</sub> Refrigeration System with a capillary. *International Journal of Refrigeration* 79, 183 - 195 (2017).
- [5] Liang Yang, Hui Li, Si-Wei Cai, Liang-Liang Shao, Chun-Lu Zhang, Minimizing COP loss from optimal high pressure correlation for transcritical CO<sub>2</sub> cycle. *Applied Thermal Engineering* 89, 656 – 662 (2015).
- [6] Y.B. Tao, Y.L. He, W.Q.Tao, Z.G.Wu, Experimental study on the performance of CO<sub>2</sub> residential air-conditioning system with an internal heat exchanger. *Energy Conversion and Management* 51, 64 – 70 (2010).
- [7] Tzong-Shing Lee, Cheng-Hao Liu, Tung-Wei Chen, Thermodynamic analysis of optimal condensing temperature of cascade-condenser in CO<sub>2</sub>/NH<sub>3</sub> cascade refrigeration systems. *International Journal of Refrigeration* 29, 1100 – 1108 (2006).
- [8] Ming Ma, Jianlin Yu, Xiao Wang, Performance evaluation and optimal configuration analysis of a CO<sub>2</sub>/NH<sub>3</sub> cascade refrigeration system with falling film evaporator–condenser. *Energy Conversion and Management* 79, 224 – 231 (2014).
- [9] S.M. Hojjat Mohammadi, Mehran Ameri, Energy and exergy performance comparison of different configurations of an absorption-two-stage compression cascade refrigeration system with carbon dioxide refrigerant. *Applied Thermal Engineering* 104, 104 – 120 (2016).
- [10] Yufei Cai, Chunling Zhu, Yanlong Jiang, Hong Shi, Modeling and calculation of open carbon dioxide refrigeration system. *Energy Conversion and Management* 89, 92 – 98 (2015).
- [11] Peihua Li, J.J.J. Chen, Stuart Norris, Flow condensation heat transfer of CO<sub>2</sub> in a horizontal tube at low temperatures. *Applied Thermal Engineering* 130, 561 – 570 (2018).
- [12] Vivek Patel, Deep Panchal, Anil Prajapati, Anurag Mudgal, Philip Davies, An efficient optimization and comparative analysis of cascade refrigeration system using NH<sub>3</sub>/CO<sub>2</sub> and C<sub>3</sub>H<sub>8</sub>/CO<sub>2</sub> refrigerant pairs. *International Journal of Refrigeration* 102, 62 – 76 (2019).

# Research on Using PSO Algorithm to Optimize Controlling of Regenerative Braking Force Distribution in Automobile

Tung Duong Tuan  
Faculty of Vehicle and Energy  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
tungdt@hcmute.edu.vn

Dung Do Van  
Faculty of Vehicle and Energy  
Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
dodzung@hcmute.edu.vn

Thinh Nguyen Truong  
Faculty of Mechanical Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
thinhnt@hcmute.edu.vn

**Abstract**— This study will analyze the optimization algorithms used in the optimization control of regenerative braking force distribution in automobile. Through these analyses, the PSO algorithm has been used to control the regenerative braking force distribution and mechanical braking force with a multi-objective function that is to maximize recovered energy and ensure braking stability. The simulation calculations are performed by MathLAB based on Toyota Hiace parameters with the traditional powertrain. Research results show that the fuel consumption rate of vehicles equipped energy recovery system after optimization can be improved from 10.49% to 24.44% depending on different driving cycles.

**Keywords**— Regenerative Braking System, Particle Swarm Optimization, Braking Force Distribution.

## I. INTRODUCTION

Braking system on automobile is a safety system. Braking is the process of transferring mechanical energy into heat at braking components. This process requires an amount of energy that the vehicle itself needs to burn the fuel to supply it. Even when the mechanical braking system wastes so much energy, it is still in use due to safety reasons. The regenerative braking system was introduced to regenerate the wasted inertial energy of the vehicle during braking phase or deceleration in order to save up fuel and increase the braking system components lifetime [1]. One of the important features of vehicles equipped with the RBS is that, in the emergency case of harsh braking, the braking moment required is more than the electric braking moment. In that case, the RBS is also activated to secure the safety. Achieving high efficiency in using the RBS is mainly focused on regenerative braking force and frictional braking force distributions. On the vehicle, the regenerative braking force distribution is illustrated by a parabola graph in fig. 1. If the reality braking force distribution curve is under the ideal one, the front wheels are braked sooner than the rear, leading to the lack of stability of the vehicle [2].

$$\frac{F_{bf}}{W_f} \geq \frac{F_{br}}{W_r} \quad (1)$$

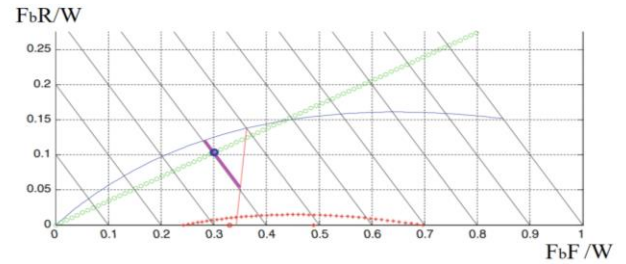


Fig. 1. Characteristic of braking force distribution

Where,  $F_{bf}$  and  $F_{br}$  are simultaneously the front and rear braking force (N);  $W_f$  and  $W_r$  is the weight of the vehicle put on the front and rear axles (N).

However, when the working point is below the ideal braking force distribution curve, most of braking force will be transmitted to the front wheels. Only a minor of it is for the rear wheels. This occasion leads to the decrease in traction of the rear wheels. In order to avoid this, one or many braking conditions are added by ECE in order to achieve the maximum front wheel braking force but in the limit (the red curve of the fig.1). based on this adjustment, the braking force distribution must meet with the following demands so that the braking intensity will be between 0.2 to 0.8 [2].

$$Z \geq 0.1 + 0.85(\mu_{ROAD} - 0.2) \quad (2)$$

In this case,  $Z$  is the braking ratio of the vehicle and  $\mu_{ROAD}$  is the traction of the road. Then, the allowed area of the braking force distribution is between the two areas in the figure 3.1. for conventional vehicle, the frictional braking system is defined as the slope of the dashed line in the fig. 1. The relationship between braking force of the front and rear wheels before stalling including the traction of the tires is ( $\mu_{ROAD}$ ) [2].

$$F_{bR} = \frac{wb - \mu_{ROAD} \cdot h}{\mu_{ROAD} \cdot h} \cdot F_{bF} - \frac{W \cdot L_b}{h} \quad (3)$$

This research is going to analyze the braking force distribution method and use the optimized control algorithm to optimize the recovered energy along with ensuring the optimal braking condition, especially the braking force distribution curve must be in the limit set by ECE.



## II. ANALYZE THE METHODS OF CONTROLLING THE BRAKING FORCE DISTRIBUTION

In order to figure the best efficiency of the regenerative and inertial braking force distribution, from the energy perspectives, there are three methods that are currently researched and applied. among the three, there is one common regulation [2].

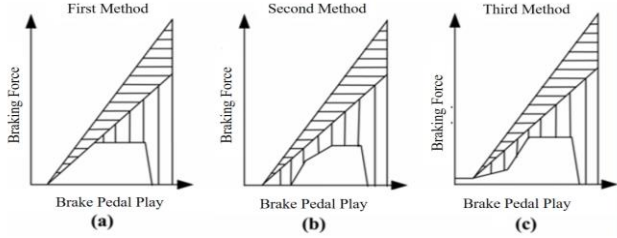


Fig. 2. Braking force distribution method graphs

### A. Maximizing the recovered energy method

This method is illustrated in fig. 2a. The regenerative moment is activated until it reach its limit. The hydraulic braking is not activated at the wheel driving the generator until its moment cannot meet with the moment required. Theoretically, this method will maximize the usage of regenerative moment and reach the highest amount of recovered energy. However, when the hydraulic braking force is needed, supplying braking oil will be decreased due to the sudden decline in pressure inside the main cylinder. This is the main reason of discomfort feeling during braking.

This method chooses the areas that points reaches out of the deceleration curve (pink curve in fig.1) and calculate regenerative energy on each point. Finally, it selects the points having the highest regenerative energy. The moment distribution is determined by the formula (4) [2].

$$T_{EM} = \begin{cases} \frac{P_{GE\_max} \times 9500}{1500}, & n_{EM} \leq 1500 \text{ r/min} \\ \frac{P_{GE\_max} \times 9500}{n_{EM}}, & n_{EM} > 1500 \text{ r/min} \end{cases} \quad (4)$$

$$\omega_b = \prod \omega = \omega_1(U) \times \omega_1(v_{SS}) \times \omega_{EM} \quad (5)$$

$$T_{EM\_reg} = \min(T_{EM}\omega_b, T_{U\_max}) \quad (6)$$

As  $\omega_1, \omega_2$  are determined:

$$\omega_1(U) = \begin{cases} 1, & 30 \leq U \leq 46 \\ -\frac{1}{2}U + 24, & 46 \leq U \leq 48 \\ 0, & 48 < U \leq 50 \end{cases} \quad (7)$$

$$\omega_2(U) = \begin{cases} 0, & v_{SS} < 10 \\ \frac{1}{20}v_{SS} - \frac{1}{2}, & 10 \leq v_{SS} \leq 30 \\ 1, & 30 < v_{SS} \leq v_{max} \end{cases} \quad (8)$$

### B. Optimizing the braking force distribution method

This method is showed by fig. 2b. When the braking process starts, only the mechanical one is activated. When the braking pressure of the front wheels is stabilized, regenerative moment will be activated. This method is ideal including for the front and rear braking forces. One working point on the parabola uses the maximum braking force, ensure safety in case the two axles are braked at the same time. If the vehicle is braked with the same traction coefficient for all wheels, when they are stalled, the force distribution is as follow:

$$F_{z1} = \frac{G}{L}(b + \phi h_g); F_{z2} = \frac{G}{L}(a - \phi h_g) \quad (9)$$

With  $\phi$  is the traction coefficient between tires and road surface. When the front and rear wheels are locked at the same time with different traction coefficient, the braking forces are determined:

$$F_{xb1} + F_{xb2} = \phi G; F_{xb1} = \phi F_{z1}; F_{xb2} = \phi F_{z2} \quad (10)$$

$$F_{xb2} = \frac{1}{2} \left[ \frac{G}{h_g} \sqrt{b^2 + \frac{4h_g L}{G} F_{xb1}} - \left( \frac{Gb}{h_g} + 2F_{xb1} \right) \right] \quad (11)$$

When the braking force is followed the ideal curve, the total braking force for the front and rear wheels is determined by the expression (11). The equation (11) is the short curve  $I_{curve}$  showing the relation between the braking force distribution between wheels when they are locked with different traction coefficient  $\phi$ . The parabola  $I_{curve}$  is illustrated in fig.3 [3], [4]

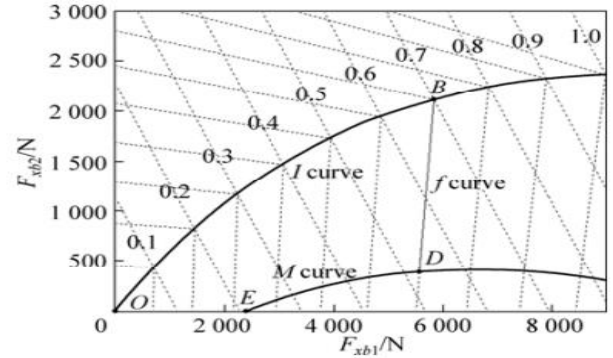


Fig. 3. Safety limit of the distribution

### C. Controlling method combination

This method is shown in fig. 2c which is the one balancing the usage of the two methods above. This method is useful in optimizing the energy recovered, pushing its to the limit while maintaining the braking feeling at the brake pedal. During the brake pedal free-play range, a small amount of moment from the motor is activated just like on the conventional vehicle. When the brake pedal is pressed hard enough, the braking pressure from the front wheels will help in supplying the total pressure while the regenerative force will reciprocate the amount that left unfilled. When the braking pressure reaches a specific value, the regenerative moment will be brought up to maximum. Until the hydraulic braking force of the front wheels requires more pressure, the oil pressure will be sent to via the boosters. Finally, the system will regain its good braking feeling, the stability in braking signal and optimization in the recovered energy.

### D. Analyzing the control algorithm

The control algorithms for regenerative braking force distribution have vital roles in recovering the energy along with stabilizing the vehicle during braking phase. There were researches applying the control algorithms such as Fuzzy Logic, PSO or the combination of the two [5]. Liuo Qin used ADVISOR software to evaluate the efficiency of the control algorithms, between the combination of PSO with Fuzzy Logic and the conventional Fuzzy Logic, in his research [5], [6].



TABLE I. THE ALGORITHMS COMPARISON

Type	Regenerative Energy[J]	Braking Energy [J]	Recovery Efficiency[%]
Fuzzy Logic	565	1697	33.29 %
Fuzzy Logic + PSO	687	1697	40.48 %

The result from this research showed that the combination of PSO and Fuzzy Logic is more efficient at the low-speed range by 7.19% compare to the conventional Fuzzy Logic. MOOP (Multi-Objective Optimization) method is executed with three objectives: regenerative energy, deceleration duration and stability during braking. These objectives are mutual. ODS (Optimization and Decision System) is determined by solving the Multi-Objective optimization problems by applying evolution algorithm.

The PSO method was succeeded in optimizing the control algorithm with the aim of advancing the energy recovery efficiency along with maintaining the vehicle stability [6]. The research area of PSO is wider than MOOP. MOOP algorithm is only suitable for autonomous vehicle (electric vehicle) as the objective of this method is to optimize the control mode and decrease the braking duration. It will be inappropriate for vehicles with driver. Therefore, the optimization using PSO will be more optimal [5], [6].

### III. OPTIMIZING THE CONTROL METHOD BY USING PARTICLE SWARM OPTIMIZATION – PSO

Designing the target function of the control unit needs to ensure 2 factors: braking efficiency and recovered energy. However, the braking force distribution is brought out by the differences between the two. The intelligent controlling algorithm is usually used in the controlling technologies in order to equalize the two. Qu Daohai concluded that braking efficiency and energy recovery are the two targets requiring optimization. The author has used Multiple Objective Particle Swarm Optimization (MOPSO) to control the regenerative braking force distribution for the Hybrid vehicle to achieve the maximum braking force [6]. Zhang Fengejiao has solved how to optimize multi-object balancing the braking stability and energy recovery efficiency based on the genetic algorithm [7]. In this research, the author also used PSO algorithm to solve the problem of optimizing multi-objects: optimize the control algorithm and braking stability.

#### A. Strategy for controlling the braking stability

Assume that all the requirements for the braking stability and energy recovery are in the suggested to be in the safety zone shown in fig. 4. The marked area OABC is the braking force distribution area that both front and rear wheel value must be inside [8].

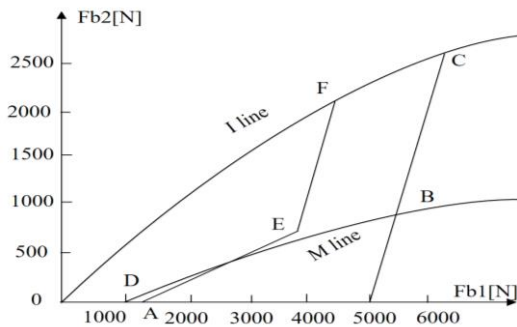


Fig. 4: Braking force distribution area's safety area

As Fb1, Fb2 are the braking forces on the front and rear axles. I curve is the ideal braking force distribution curve. M line is adjustment from ECE-R13. The line f is brake characteristics when the front wheels are locked. D is the point that match with the braking intensity 0 value. E is the maximum regenerative braking point. DE line is tangent to M line. EF line appears when the front wheels are locked.

#### B. Model for optimizing the strategy for controlling regenerative braking force.

Braking force distribution includes two factors. The first one is the distribution between the front and rear wheels. The second is the distribution between regenerative and hydraulic braking force. The first factor mainly affects the vehicle stability during braking. However, the latter will directly affect the energy recovery efficiency. Therefore, this research has taken the two above reasons as the main targets to optimize.

Choose the variables: based on the above analysis, the two target functions are closely related to the braking force distribution. Optimizing the control strategy is mostly about optimizing braking force distribution. Therefore, the X variables must be [9]:

$$X = [F_{b1}, F_{b2}, F_{reg}] \quad (12)$$

As  $F_{b1}$ ,  $F_{b2}$  and  $F_{reg}$  represent the braking force of the front, rear and regenerative braking force simultaneously

Establish the functions: the stability when braking is chosen as the first target to be studied. The chosen traction coefficient is used to simulate the condition of the road surface [9], traction coefficient and braking intensity. The more it closes to reality, the more reasonable the distribution is. Therefore, the target functions of braking stability  $Y_1$  is established by the following formula [9]:

$$\text{Min}Y_1 = \sqrt{(\varphi_1 - z) + (\varphi_2 - z)} \quad (13)$$

As  $\varphi_1$ ,  $\varphi_2$  are correspondent to traction coefficient of the front and rear wheels;  $z$  is the braking intensity. A part of the energy transmitted from the wheels to the engine through the powertrain is transformed into electricity and is stored in the battery at the end of the braking phase. Therefore, the efficiency in energy storing of the battery is chosen as the target function  $Y_2$  [9].

$$\text{Max}Y_2 = F_m V \eta_m \eta_b \eta_{tl} \quad (14)$$

As  $\eta_m$ ,  $\eta_b$  và  $\eta_{tl}$  are the motor efficiency, battery efficiency and powertrain efficiency.  $V$  is the current vehicle velocity. The linear progression method is used to change the multi-objects problem into a single-object one, so the Y function is [9]:

$$\text{Min}Y = k_1 Y_1 - k_2 Y_2 \quad (15)$$

As  $k_1$  and  $k_2$  are weight values of two functions. These values are calculated based on the driver's control.

Establish the boundary conditions: During braking process, the regenerative braking force is not only limited by the torsional moment from the generator but also the charging ability of the battery and the front wheels braking force. In addition, to ensure the safety zone while braking, the distribution must also be in the safety zone. Therefore, the boundary conditions for the functions are [9]:

$$F_{b1}, F_{b2} \in P_{OABC}; T_m \leq T_{mt};$$

$$F_m V \eta_m \eta_b \eta_{tl} \leq P_{\max}; F_m \leq F_{b1} \quad (16)$$

As  $T_m$  is the regenerative moment.  $T_{mt}$  is the torsional moment from motor/generator.  $P_{\max}$  is the maximum recharged energy from battery.

### C. Apply the PSO algorithm.

When the vehicle speed and braking intensity change, the target functions and boundary conditions are changed as well. Therefore, it is almost impossible to solve this problem in a conventional way. The intelligent optimizing algorithm are usually used to solve these problems and the PSO (Particle Swarm Optimization) turns out to be a simple and effective way.

At the beginning, PSO initializes a group of particles in the possible-solution space, each particle will be one potential optimizing solution. After that, the algorithm will calculate and find out the optimum value. The particles are update continuously by observing two factors,  $p_{best}$  and  $g_{best}$  based on these formulas [9]:

$$\begin{aligned} v_d^i(k+1) &= w v_d^i(k) + c_1 \cdot r_1 \cdot [p_{best_i}(k) - x_d^i(k)] \\ &+ c_2 \cdot r_2 \cdot [g_{best_i}(k) - x_d^i(k)] \\ v_d^i(k+1) &= v_d^{\max}, \text{ if } v_d^i(k+1) > v_d^{\max}; \\ v_d^i(k+1) &= v_d^{\min}, \text{ if } v_d^i(k+1) < v_d^{\min} \\ x_d^i(k+1) &= x_d^i(k) + v_d^i(k+1) \\ x_d^i(k+1) &= x_d^{\max}, \text{ if } x_d^i(k+1) > x_d^{\max}; \\ x_d^i(k+1) &= x_d^{\min}, \text{ if } x_d^i(k+1) < x_d^{\min} \end{aligned} \quad (17)$$

When:  $v_d^i(k)$ ,  $x_d^i(k)$  is the velocity and the partial position of the  $k$  generation and  $i$  particle in a swarm that its size is  $d$ ;  $v_d^{\max}$ ,  $x_d^{\max}$  are the velocity of the position of the maximum parts in the swarm  $d$ .

### D. Simulation and analyzation from the result

TABLE II. SIMULATION VEHICLE PARAMETERS (TOYOTA HIACE)

Parameters	Value
Wheel base (mm)	2570
Width (mm)	1430
Ground clearance (mm)	182
Weight (kg)	1905
Maximum power (kW/rpm)	74/5400
Maximum Torque (Nm/rpm)	165/2600
Ratio at 1 <sup>st</sup> gear	4.452
Ratio at 2 <sup>nd</sup> gear	2.619
Ratio at 3 <sup>rd</sup> gear	1.517
Ratio at 4 <sup>th</sup> gear	1.000
Ratio at 5 <sup>th</sup> gear	0.854
Reverse ratio	4.472
Final drive ratio	4.3
Frontal area [m <sup>2</sup> ]	2.325
Wheel radius [m]	0.33

TABLE III. REGENERATIVE BRAKING ENERGY RECOVERY ASSEMBLY PARAMETERS

Parts	Dimension parameters		
	Teeth	Pitch [mm]	Dimension [mm]
Front ring gear	65	2.5	162.5
Front sun gear	27	2.5	67.5
Front planetary gear	19	2.5	47.5
Rear ring gear	73	2.5	182.5
Rear sun gear	39	2.5	97.5
Rear planetary gear	17	2.5	42.5
Flywheel	Dimension: 220x30 mm. Weight: 2.7 kg. Material: Iron. Inertia moment: 0.04 kg.m <sup>2</sup> .		
Generator	1.5 hp		

Simulation process description: When the simulation starts, all the standard driving cycles will be uploaded onto the model. The PID controller will control the provided acceleration and deceleration process of the cycle. When there is signal for deceleration or braking, the required braking force will be calculated. If the required braking force is higher than the regenerative braking force, then the maximum regenerative braking force will be applied to achieve maximum recovered energy and the remain will be the mechanical one. On the contrary, if the required braking force is lower or equal to the regenerative braking force, the mechanical brake will not work and only the RBS will be activated. The simulation model combined MathLAB Simulink and CarSim has the following flowchart in Fig. 5.

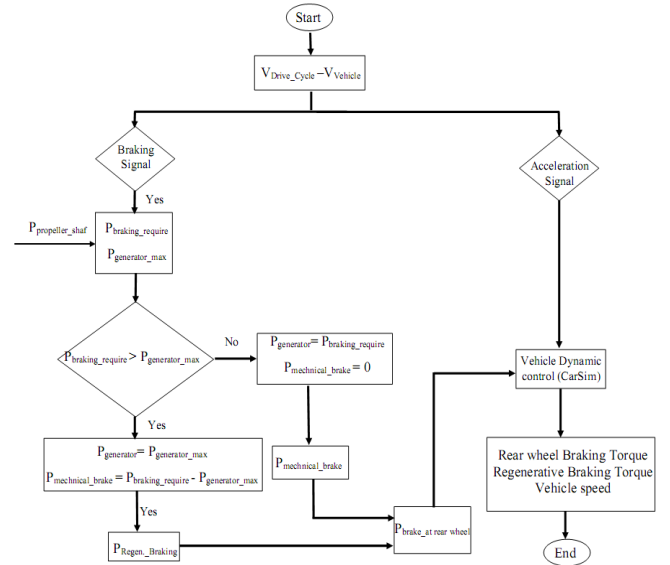


Fig. 5. Regenerative braking force distribution flowchart

The optimization solution using the PSO is simulated by MathLAB. The parameters of PSO is established with 600 swarms. The maximum loops are 100. Both  $c_1$  and  $c_2$  are set by 2. PSO method with the linear decreased inertial mass is applied in finding the points having optimum regenerative braking force distribution. All the optimal solutions are shown in Fig. 6.

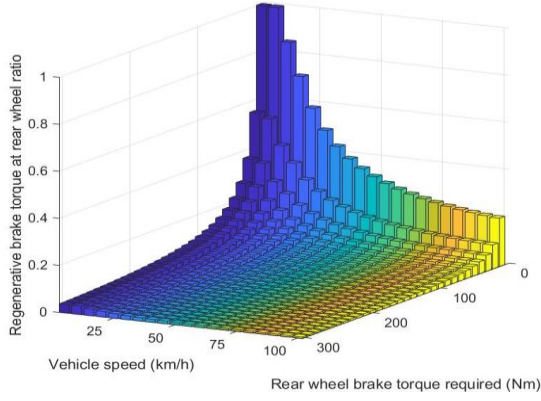


Fig. 6. Optimal braking force distribution map

Based on the fig. 6, during the low velocity range, if the rear wheel braking force is required (the shaft driving the energy recovery system), the controller will allow to maximize the regenerative braking force by adjusting the SOC charging coefficient. The usage of the regenerative braking force will increase which leads to a decrease in using mechanical one. When the velocity increases, the required braking force is also increased causing the insufficiency if only the regenerative braking force is used. Therefore, the controller will activate the mechanical braking mechanism to maintain stability when braking. The usage ratio of the regenerative braking force is decreased.

In term of safety and vehicle stability in braking phase, the simulation model is set up based on the optimal solution for the braking force distribution. The vehicle will start braking at different points with various velocity values, ranging from 5km/h to 100km/h and based on the standard driving cycle with the braking density increasing from 0 to 0.7.

The higher the traction coefficient is, the higher the braking density will be. This allows the braking force distribution to be more reasonable. The optimal front and rear braking force distribution can both meet with the braking force distribution regulations. However, their tendency curves are even rational. From this result, we can conclude that optimal control strategy can increase braking stability efficiency.

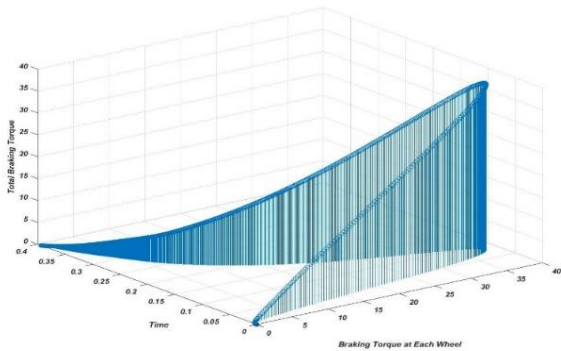


Fig. 7. Braking force distribution at wheels

#### E. Calculate the recovered energy and fuel consumption rate

So, the total fuel consumption for one testing cycle is:

$$Q_t = \frac{A_t g_{etb}}{36.10^5} + \frac{G_{xx}(v_1 - v_2)}{3600 j_{tb}} \quad (18)$$

If the distance travelled by the vehicle is can be determined in case the acceleration is  $S_i$  and the inertial movement is  $S_{li}$ , the fuel consumption on one unit of distance can be calculated as follow:

$$Q_{st} = \frac{100Q_t}{(S_j + S_{li})\rho_n} \quad (19)$$

Therefore, the fuel consumption is determined by the expression:

$$G_T = \frac{Q\rho_n}{t} \quad (20)$$

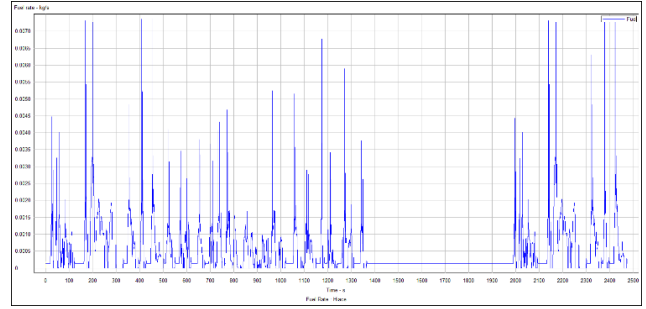


Fig. 8. The graph showing the rate of fuel consumption based on each driving cycles

In order to evaluate the fuel consumption of the engine, the useable fuel consumption  $g_e$  is used:

$$g_e = \frac{G_T}{N_e} = \frac{Q\rho_n}{N_e t}; q_d = \frac{100Q}{S^x} \quad (21)$$

As:  $N_e$  is the actual power (kW). Through the engine experiment and calculations, the graph illustrating the relationship between engine power  $N_e$  and the fuel consumption with the crankshaft RPM  $N_e = f(n_e)$  and  $g_e = f(n_e)$ . Then, the fuel consumption will be:

$$q_d = \frac{100g_e N_e t}{S^x \rho_n} = \frac{100g_e N_e}{v \rho_n} \quad (22)$$

$\rho_n$  – fuel density (kg/l). During the simulation, the vehicle is driven with the differential velocity. Therefore, the fuel consumption for unstable movement is:

$$q_d = \frac{(P_{\psi} + P_{\omega} \pm P_j) 0.36 g_e}{\rho_n \eta_t} \quad (23)$$

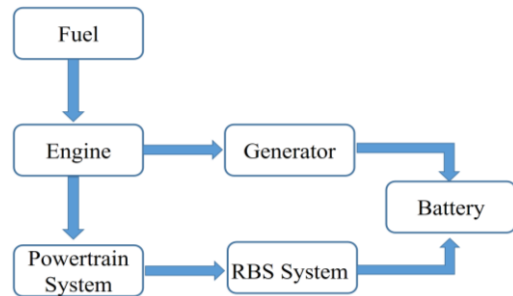


Fig. 9. Energy flow on the vehicle

Engine efficiency:  $\eta_e = 0.2 - 0.35$

Powertrain efficiency:  $\eta_d = 0.95 - 0.96$

Generator efficiency:  $\eta_p = 0.4 - 0.65$

Battery efficiency:  $\eta_a = 0.75 - 0.9$

PBS's Planetary gearset efficiency:  $\eta_b = 0.95 - 0.97$

Total work efficiency:  $\eta_t = \eta_e \cdot \eta_d \cdot \eta_p \cdot \eta_a \cdot \eta_b = 0.054 - 0.190$ .

### F. Result and Discussion

Based on the simulation result, the recovered energy and fuel consumption rate are calculated based on each driving cycle in the table IV

TABLE IV. THE SPECIFICATIONS AFTER OPTIMAL CONTROL STRATEGY IS APPLIED

Driving Cycles		FTP-75	NEDC	EUDC	ECE 15
Test Duration [s]		3748	1180	400	195
RBS working duration [s]	Before	1145	238	94	36
	After	1455	312	124	40
Working duration [%]	Before	30.5	20.2	23.5	18.5
	After	38.7	26.5	31.1	20.7
Total energy [kJ]	Before	18038.4	2478.1	1745.5	209.1
	After	22915.5	3252.9	2309.4	233.6
	Before	9.82	9.53	9.42	8.48
	After	7.73	7.26	7.12	7.59

The simulation and calculation results shown that after applying the PSO control algorithm the working duration of the energy recovery unit increases depend on each driving cycle. Therefore, the amount of energy recovered also increases. This energy will recharge the battery for other accessories usage on the vehicle. Thus, loads putting on the generator will be decreased which directly lead to a reduction in the fuel consumption.

The fuel consumption rate of the vehicle equipped with the regenerative braking system can be improved from 10,49% to 24,44% depend on which driving cycle is being applied. This is achieved by optimizing the regenerative braking force and mechanical braking force distribution which advance the amount of energy recovered while maintain the vehicle stability. The traction coefficient and the braking force

distributed area are still in the allowed limit of the ECE standard on regenerative braking force distribution.

### ACKNOWLEDGMENT

This study is a successful result with the tremendous support from scientists in the Automotive, Mechanical engineering field at the Ho Chi Minh City University of Technology and Education. Authors would extend my thanks to all the University Managing Board, Faculty of Vehicle and Energy Engineering, Faculty of Mechanical Engineering for facilitating in our research.

### REFERENCES

- [1] S.J.Clegg (1996) A Review of Regenerative Brake System. Institute of Transport Studies, University of Leeds.
- [2] Gao H, Yimin Gao Y and Ehsani M. A neural network based SRM drive control strategy for regenerative braking in EV and HEV. In: IEEE international electric machines and drives conference, Cambridge, MA, USA, 17–20 June 2001, pp.571–575. New York: IEEE.
- [3] Aoki Y, Suzuki K, Nakano H, et al. Development of hydraulic servo brake system for cooperative control with regenerative brake. SAE paper 2007-01-0868, 2007.
- [4] Bradley Glenn, Gregory Washington, Giorgio Rizzoni. Operation and control strategies for hybrid electric automobiles. SAE2000-01-1537, 2000.
- [5] Liao Qin “Particle Swarm Optimization Algorithm for Regenerative Braking Fuzzy Control of Electric Vehicle” International Conference on Information Sciences, Machinery, Materials and Energy (ICISMME 2015)
- [6] D. H. Qu, Research of regenerative braking energy feedback control strategy for electric vehicle, Hunan: Hunan University, 2014.
- [7] F. J. Zhang, M. X. Wei, Multi-objective optimization of the control strategy of electric vehicle electro-hydraulic composite braking system with genetic algorithm, Adv. Mech. Eng. 7(2) (2015).
- [8] H. L. Liu, X. P. Dong, B. L. Zhang, “Study on control strategy of regenerative braking for electric vehicle”, J. Hefei Univ. Technol. (2009) 108-120.
- [9] Guo Zhijun, Yue Dongdong and Wu Jingbo “Optimization of Regenerative Braking Control Strategy for Pure Electric Vehicle”

# Design Depth Controller for Hybrid Autonomous Underwater Vehicle using Backstepping Approach

Nguyen-Nhut-Thanh Pham  
Ho Chi Minh City University of  
Technology, VNUHCM  
Ho Chi Minh City, Vietnam Vietnam  
pnnthanh.beray@gmail.com

Ngoc-Huy Tran✉  
Ho Chi Minh City University of  
Technology, VNUHCM  
Ho Chi Minh City, Vietnam Vietnam  
tnhuy@hcmut.edu.vn

Thien-Phuong Ton  
Ho Chi Minh City University of  
Technology, VNUHCM  
Ho Chi Minh City, Vietnam Vietnam  
tonphuong@hcmut.edu.vn

Thien-Phuc Tran  
Ho Chi Minh City University of  
Technology, VNUHCM  
Ho Chi Minh City, Vietnam  
tphuc.rectie@hcmut.edu.vn

**Abstract**— This paper presents a study on controller design for the hybrid AUV in the vertical plane using the backstepping approach. Firstly, the six degrees of freedom (6-DOF) nonlinear kinematic and dynamic equations of motion for the hybrid AUV is established, and the operating mechanisms of the hybrid AUV are described. Subsequently, the reduced order mathematical model for depth control will be drawn from this 6-DOF model by decoupling and linearizing. Next, the depth controller is designed by the backstepping technique and its stability will be guaranteed by using Lyapunov's theorem. Finally, the results will be presented through the MATLAB/SIMULINK simulation which will prove the effectiveness and feasibility of the proposed method.

**Keywords**— AUV, Depth control, Backstepping technique

## I. INTRODUCTION

Nowadays, autonomous vehicles are gradually replacing people in many different fields, especially in tasks in which potential risks or environments where humans cannot reach. In deep oceanic missions, autonomous underwater vehicles (AUVs) have become one of the best solutions because of its safety, effectiveness, economic savings. These missions involve data collection and environmental sampling for hydrological research or oceanographic and biological surveys, pipeline inspection, seafloor mapping, offshore oil and gas exploration and exploitation, and mine countermeasure,...([1-7]). To improve the performance of AUV to fulfill the aforementioned missions, the research and development of control algorithms for AUV are indispensable.

During the past few years, many algorithms of AUV motion control are reported. In most articles, AUV's motion control problem will be divided into three sub-problems to simplify and facilitate the construction of the controller. These sub-problems include steering, diving, and speed control. In particular, the diving control problem has attracted a lot of attention, thereby several control algorithms have been studied with different focuses to deal with the highly nonlinear and strongly coupled system. For instance, some algorithms that have been utilized to build depth controller are PID ([8-10]), Sliding Mode Control (SMC) ([11-13]), Backstepping technique ([14], [12], [15]), Dynamic Inversion (DI) ([16]),

$H_\infty$  control ([2], [15]),  $H_2$  control ([2], [17]), Adaptive control ([18])...

In previously published articles, the model commonly used to design the depth controller and simulation is the 3-DOF model, where the sway velocity, yaw angle, and roll angle are approximately zero. The use of a reduced-order model is necessary because it decreases the complexity of the control laws and makes it easy to apply them in practice [19]. However, the designed controller can be unstable due to the oscillation of the roll angle during control [20], thereby the simulation results must still be verified on the 6-DOF AUV model to guarantee the objectivity and feasibility of the designed controller. Prompted by these observations, this paper presents the controller design for the depth-tracking problem of a hybrid AUV using the backstepping technique. A reduced-order model for depth-plane is first established by linearizing and decoupling the 6-DOF AUV model. The control laws are designed to guarantee the depth tracking error converge to an arbitrarily small neighborhood of zero. To verify that the proposed controller can operate effectively and sustainably with the oscillation of roll angle during control, numerical simulations are carried out with the 6-DOF AUV model.

The remainder of this paper is organized as follows. Section 2 briefly describes the 6-DOF AUV models, the operating mechanisms of the hybrid AUV, the linearized depth-plane model, and formulates control problem. Section 3 presents the controller design for the depth-plane motion of the AUV by using the backstepping technique. Section 4 validates and discusses the control performance of the proposed controller through numerous numerical simulations. Finally, Section 5 summarizes the main contributions of this paper.

## II. PROBLEM FORMULATION

In this paper, AUV2000 mentioned in Fig. 1 [13] will be chosen to be the object of study and controller design. Although AUV2000 can operate in both AUV modes (using thruster) and Glider mode (without using the thruster), this paper only focuses on developing the depth controller in AUV mode.



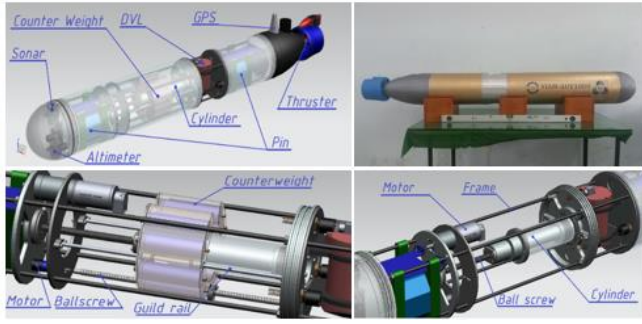


Fig. 1. Overview of AUV2000 system.

### A. AUV modelling

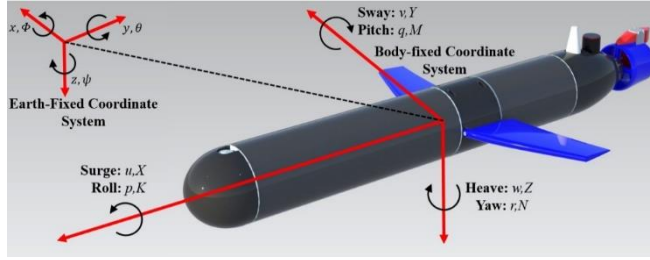


Fig. 2. AUV2000 Body-Fixed and Earth-Fixed Coordinate Systems.

According to [21], the 6-DOF nonlinear equations of motion of the AUV can be described in two reference frames are earth-fixed frame {e} and the body-fixed frame {b} as shown in Fig. 2. The kinematics of AUV can be expressed as:

$$\begin{cases} \dot{x} = u \cos \psi \cos \theta + v (-\sin \psi \cos \phi + \cos \psi \sin \theta \sin \phi) \\ \quad + w (\sin \psi \sin \phi + \cos \psi \sin \theta \cos \phi) \\ \dot{y} = u \sin \psi \cos \theta + v (\cos \psi \cos \phi + \sin \psi \sin \theta \sin \phi) \\ \quad + w (-\cos \psi \sin \phi + \sin \psi \sin \theta \cos \phi) \\ \dot{z} = -u \sin \theta + v \cos \theta \sin \phi + w \cos \theta \cos \phi \\ \dot{\phi} = p + q \sin \phi \tan \theta + r \cos \phi \tan \theta \\ \dot{\theta} = q \cos \phi - r \sin \phi \\ \dot{\psi} = q \frac{\sin \phi}{\cos \theta} + r \frac{\cos \phi}{\cos \theta} \end{cases} \quad (1)$$

Beside, following to [13], the 6-DOF nonlinear dynamic equations of motion of AUV2000 can be expressed as

$$\begin{aligned} (m - X_u) \dot{u} - m y_G \dot{r} + m z_G \dot{q} \\ = (Z_w - m) w q + (Z_{\dot{q}} + m x_G) q^2 + (m - Y_v) v r \\ + (m x_G - Y_r) r^2 - m y_G p q - m z_G p r \\ - (W - B) \sin \theta + X_{u|u|} u |u| + X_{prop} \end{aligned} \quad (2a)$$

$$\begin{aligned} (m - Y_v) \dot{v} - m z_G \dot{p} + (m x_G - Y_r) \dot{r} \\ = (X_u - m) u r + (m - Z_w) w p - (m x_G + Z_{\dot{q}}) p q \\ + m y_G (p^2 + r^2) - m z_G q r + Y_{v|v|} v |v| \\ + Y_{r|r|} r |r| + (W - B) \cos \theta \sin \phi \\ + Y_{uvl} u v + Y_R \end{aligned} \quad (2b)$$

$$\begin{aligned} (m - Z_w) \dot{w} - (m x_G + Z_{\dot{q}}) \dot{q} + m y_G \dot{p} \\ = (m - X_u + Z_{uqf}) u q + (Y_v - m) v p - m y_G r q \\ + m z_G (p^2 + q^2) + (Y_r - m x_G) r p \\ + Z_{w|w|} w |w| + (W - B) \cos \theta \cos \phi \\ + Z_{q|q|} q |q| + (Z_{uvl} + Z_{uwf}) u w \end{aligned} \quad (2c)$$

$$\begin{aligned} (I_{xx} - K_p) \dot{p} + m y_G \dot{w} - m z_G \dot{v} \\ = (Z_w - Y_v) w v + (Z_{\dot{q}} + Y_r) v q - (Z_{\dot{q}} + Y_r) w r \\ + m z_G (u r - w p) - m y_G (v p - u q) \\ + (N_r - M_{\dot{q}} - I_{zz} + I_{yy}) q r + K_{p|p|} p |p| \\ + (y_G W - y_B B) \cos \theta \cos \phi \\ - (z_G W - z_B B) \cos \theta \sin \phi + K_{prop} \end{aligned} \quad (2d)$$

$$\begin{aligned} (I_{yy} - M_{\dot{q}}) \dot{q} - (m x_G + Z_{\dot{q}}) \dot{w} + m z_G \dot{u} \\ = (X_u - Z_w + M_{uvl} + M_{uwf}) u w + M_{w|w|} w |w| \\ + (K_p - N_r - I_{xx} + I_{zz}) r p + M_{q|q|} q |q| \\ + (M_{uqf} - m x_G - Z_{\dot{q}}) u q \\ - m z_G (w q - v r) + (m x_G - Y_r) v p \\ - (z_G W - z_B B) \sin \theta \\ - (x_G W - x_B B) \cos \theta \cos \phi \end{aligned} \quad (2e)$$

$$\begin{aligned} (I_{zz} - N_r) \dot{r} + (m x_G - N_v) \dot{v} - m y_G \dot{u} \\ = (Y_v - X_u + N_{uvl}) u v + (m x_G + Z_{\dot{q}}) w p \\ + (M_{\dot{q}} - K_p + I_{xx} - I_{yy}) p q + N_R \\ + (Y_r - m x_G) u r + m y_G (w q - v r) \\ + (x_G W - x_B B) \cos \theta \sin \phi + N_{r|r|} r |r| \\ + (y_G W - y_B B) \sin \theta + N_{v|v|} v |v| \end{aligned} \quad (2f)$$

in which  $m$  is the vehicle mass,  $(x_G, y_G, z_G)$  is the center of gravity,  $(x_B, y_B, z_B)$  is the center of buoyancy, and  $I_{xx}, I_{yy}, I_{zz}$  respectively are inertial moments about the  $x, y, z$  axes. The parameter  $X_{(\cdot)}, Y_{(\cdot)}, Z_{(\cdot)}, K_{(\cdot)}, M_{(\cdot)}, N_{(\cdot)}$  are the added mass, damping and lift terms.  $X_{prop}, K_{prop}, Y_R$ , and  $N_R$  are the force and moment made by propeller and rudder, respectively.

### B. The specific characteristics of AUV-2000

AUV2000 is a new platform in which the counterweight and ballast system allow AUV2000 to change the center of gravity and the buoyancy, respectively. Therefore, the pitch angle of AUV2000 can be controlled by the counterweight system instead of using the stern plane as conventional types. This considerably improves the performance of the controller. In this paper, for simplicity, we will take advantage of the formulas related to the counterweight and ballast system of AUV2000 that have been built in [13] as follows.

### 1) The buoyancy center and buoyancy of AUV2000:

This system includes a piston and cylinder that allows the AUV2000 to be able to draw and release the water, thereby enabling the AUV to change the center of buoyancy and buoyancy.

Let  $V_{fix}$  is the volume occupied by AUV when the ballast system draws full of water,  $l_{fix}$  is the distance from the center of buoyancy to the origin of {b},  $V_{var}$  is the volume of water that the cylinder emits, then the formula for determining the buoyancy center and the buoyancy of AUV can be respectively expressed as

$$\vec{r}_{cb} = [x_B, 0, 0]^T = \left[ \frac{V_{var} \frac{l_c + x_p}{2} + V_{fix} l_{fix}}{V_{var} + V_{fix}}, 0, 0 \right]^T \quad (3)$$

$$B = \rho g (V_{fix} + V_{var}) \quad (4)$$

where  $l_c$  is the length of cylinder,  $x_p$  is the piston position in {b},  $\rho$  is the density of fluid, and  $g$  is the gravitational acceleration.

### 2) The center of gravity and weight of AUV2000:

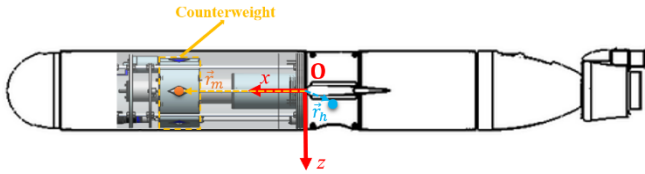


Fig. 3. Structure of the counterweight system on AUV2000.

This system is composed of a battery block (counterweight) capable of moving along the x-axis of the AUV2000 (Fig. 3), thereby making it possible to change the vehicle's center of gravity by changing the counterweight position.

Let  $m_h$  is the total weight of AUV without counterweight,  $\vec{r}_h$  is the mass center of  $m_h$  in {b}, and  $\vec{r}_m = [x_m, 0, 0]^T$  is the mass center of the counterweight  $m_m$  in {b} ( $x_m$  is the counterweight position). Then the center of gravity and the weight of AUV can be respectively expressed as

$$\vec{r}_{cg} = [x_G, y_G, z_G]^T = \frac{m_h \vec{r}_h + m_m \vec{r}_m}{m_h + m_m} \quad (5)$$

$$W = mg = (m_h + m_m)g \quad (6)$$

### C. The linearized depth-plane model

AUV is a highly nonlinear system, so for the convenience and simplicity of designing a depth controller, the use of the reduced-order model is inevitable [8], [17]. Besides, this paper only considers the motion of the AUV in the vertical plane so we can utilize the reduced kinematic equations in [18] and the simplified pitch dynamic in [13]. Then we get the 3-DOF depth-plane model for AUV2000 as follows

$$\begin{cases} \dot{z} = -u\theta \\ \dot{\theta} = q \\ \dot{q} = f_q + g_q x_G \end{cases} \quad (7)$$

where

$$f_q = \frac{(X_u - Z_w + M_{uw} + M_{uwf})uw + (M_{uf} - Z_q)uq - mz_G wq + M_{w|w}|w| + M_{q|q}|q| - (z_G W - z_B B) \sin \theta + x_B B \cos \theta}{I_{yy} - M_q} \quad (8)$$

$$g_q = -\frac{muq + W \cos \theta}{I_{yy} - M_q} \quad (9)$$

### D. Control Objective

The control object in this paper is to develop a control law  $x_G$  in (7) to force AUV2000 to converge and stabilize on the desired depth, regardless of roll angle effect. Additionally, the developed control law must demonstrate effectiveness and feasibility when applied to the 6-DOF AUV model through various numerical simulations.

### III. CONTROLLER DESIGN

In this section, the backstepping technique is utilized to deduce the control law for tracking the desired depth of AUV2000 in the vertical plane. At first, we denote the desired depth is  $z_d$  and define the depth tracking error as:

$$e_1 = z - z_d \quad (10)$$

Creating the first control Lyapunov function as follows

$$V_1 = \frac{1}{2} e_1^2 \quad (11)$$

Differentiating the variable along the (10) is obtained as

$$\dot{V}_1 = e_1 \dot{e}_1 = e_1 (-u\theta - \dot{z}_d) \quad (12)$$

The first virtual control input and an error variable for tracking the pitch angle  $\theta$  are defined as

$$\theta_d = \frac{k_1 e_1 - \dot{z}_d}{u}, \quad k_1 > 0 \quad (13)$$

$$e_2 = \theta - \theta_d \quad (14)$$

then

$$\begin{aligned} \dot{V}_1 &= e_1 (-u(e_2 + \theta_d) - \dot{z}_d) \\ &= e_1 \left( -ue_2 - u \frac{k_1 e_1 - \dot{z}_d}{u} - \dot{z}_d \right) \\ &= -k_1 e_1^2 - ue_1 e_2 \end{aligned} \quad (15)$$

Subsequently, the second Lyapunov function is defined

$$V_2 = V_1 + \frac{1}{2} e_2^2 \quad (16)$$

whose time derivative is

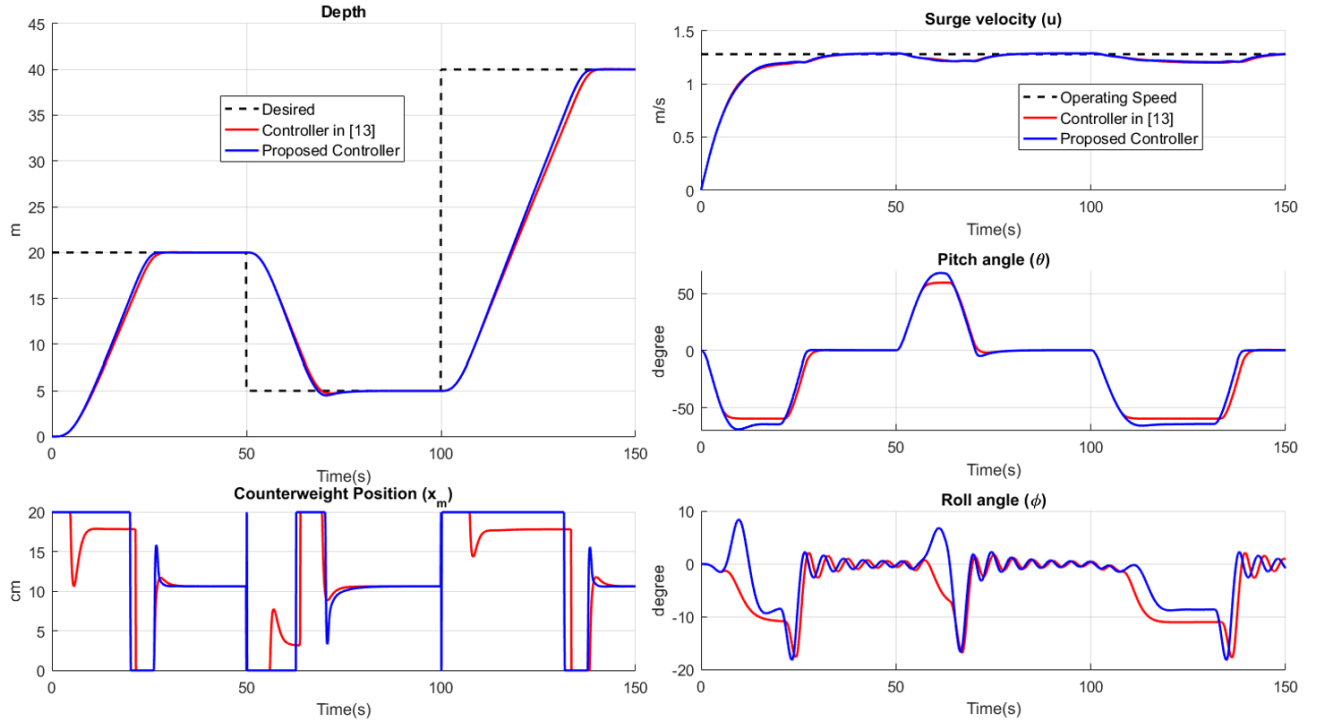


Fig. 4. Simulation results of tracking the level depth path of the controller in [13] (red solid line) and proposed controller in this paper (blue solid line).

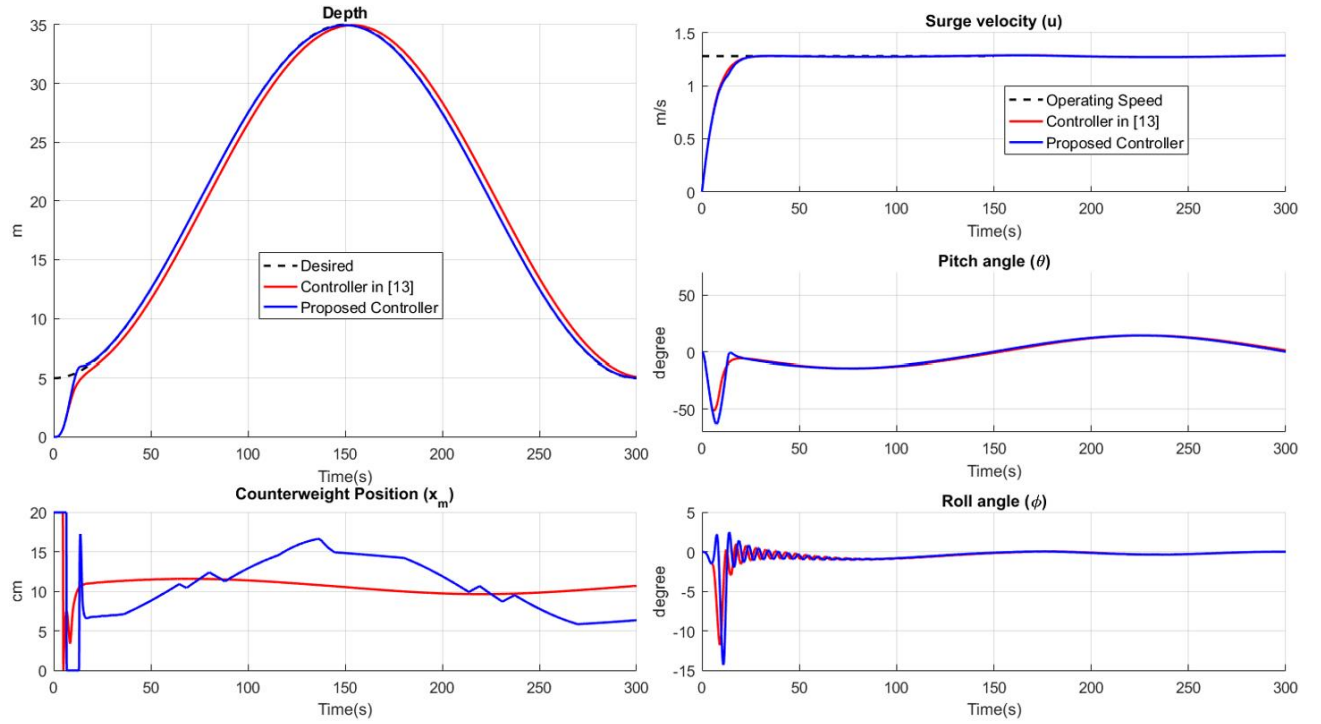


Fig. 5. Simulation results of tracking the sinusoidal depth path of the controller in [13] (red solid line) and proposed controller in this paper (blue solid line).

$$\begin{aligned}
 \dot{V}_2 &= \dot{V}_1 + e_2 \dot{e}_2 \\
 &= -k_1 e_1^2 - u e_1 e_2 + e_2 (q - \dot{\theta}_d) \\
 &= -k_1 e_1^2 + e_2 (q - \dot{\theta}_d - u e_1)
 \end{aligned} \quad (17)$$

$$q_d = -k_2 e_2 + \dot{\theta}_d + u e_1, \quad k_2 > 0 \quad (18)$$

$$e_3 = q - q_d \quad (19)$$

Substitute (18) and (19) in (17) yields

The second virtual control input and the tracking error variable of pitch rate  $q$  are selected as

$$\begin{aligned}
\dot{V}_2 &= -k_1 e_1^2 + e_2 (e_3 + q_d - \dot{\theta}_d - u e_1) \\
&= -k_1 e_1^2 + e_2 (e_3 - k_2 e_2 + \dot{\theta}_d + u e_1 - \dot{\theta}_d - u e_1) \\
&= -k_1 e_1^2 - k_2 e_2^2 + e_2 e_3
\end{aligned} \quad (20)$$

Considering the third Lyapunov function

$$V_3 = V_2 + \frac{1}{2} e_3^2 \quad (21)$$

and its derivative is

$$\begin{aligned}
\dot{V}_3 &= \dot{V}_2 + e_3 \dot{e}_3 \\
&= -k_1 e_1^2 - k_2 e_2^2 + e_2 e_3 + e_3 (f_q + g_q x_G - \dot{q}_d) \\
&= -k_1 e_1^2 - k_2 e_2^2 + e_3 (f_q + g_q x_G - \dot{q}_d + e_2)
\end{aligned} \quad (22)$$

Now, the control law is chosen as follows

$$x_G = \frac{-k_3 e_3 - f_q + \dot{q}_d - e_2}{g_q}, k_3 > 0 \quad (23)$$

Then we obtain

$$\dot{V}_3 = -k_1 e_1^2 - k_2 e_2^2 - k_3 e_3^2 < 0 \quad (24)$$

It can be concluded from (24) that the tracking errors  $[e_1, e_2, e_3]$  converge to the origin by Lyapunov's theorem.

#### IV. SIMULATION RESULTS

In order to verify the effectiveness of proposed control law, numerical simulations are conducted with the AUV 6-DOF model in this section. In practical case, for simplicity, researchers usually consider that the thrust of propeller is fixed, therefore the surge velocity  $u$  remains close to the operating speed  $U_0$ . In this paper, the operating speed  $U_0 = 1.28 \text{ m/s}$  is selected for the study. Besides, this study is divided into two cases. Case 1 is carried out to verify the performance of the proposed depth controller compare to other conventional controllers ([9], [13], [17]). Case 2 is implemented to illustrate the tracking ability of proposed control law. The control gains selected for all case simulations are  $k_1 = 1.2, k_2 = 4.6, k_3 = 0.12$ .

##### A. Case 1: Tracking the level depth path

The simulation results from Fig. 4 show that the proposed controller operate very well, which can help the vehicle to achieve the desired depth despite being affected by the roll angle (maximum 20 degrees). Besides, the pitch angle during operation is always within the allowable limits of the vehicle (less than 70 degrees), and the surge velocity remains close to the operating speed  $U_0$ . Moreover, the control performance comparison in Fig. 4 also demonstrated the superiority of the proposed controller, namely that the proposed controller had a faster response time, less overshoot than the controller in [13].

##### B. Case 2: Tracking the sinusoidal depth path

Fig. 5 shows simulation results of tracking the sinusoidal desired path  $z_d = 20 - 15 \cos(\pi t / 150)$ . From this result, we can see that the controller in [13] is incapable of making the vehicle follow the time-varying desired path. However, the tracking ability of the proposed controller is demonstrated that

not only forces the AUV to the desired path quickly but also stabilizes the vehicle on the desired path. Besides, the dynamic responses of the vehicle are smooth because the desired path doesn't suddenly change as case 1.

From all simulation results and aforementioned analyses, it can be concluded that the proposed controller can be tracking not only straight lines but also curve paths. The control signal during the maneuvering is smooth and applicable in practice. Besides, the performance of the controller is maintained regardless of the roll angle effect. Therefore, the effectiveness and feasibility of the proposed method have been proved.

#### V. CONCLUSION

This paper has developed an alternative method for depth tracking control of a hybrid AUV. In this paper, the backstepping technique has been utilized to build the depth tracking controller for AUV2000 in the vertical plane. The evaluation simulation results are applied to the 6-DOF model to investigate the effect of the roll angle on the controller's stability and quality. Through the simulation results, the effectiveness and feasibility of the proposed method are proved, it not only helps the vehicle to be able to achieve both the straight and curved desired depth path but also has a smooth control signal feasible for practice applications.

#### ACKNOWLEDGMENT

This research is supported by National Key Lab. of Digital Control and System Engineering (DCSELAB), HCMUT; Laboratory of Advance Design and Manufacturing Processes and funded by Vietnam National University Ho Chi Minh city (VNU-HCM) under grant number B2018-20b-01.

#### REFERENCES

- [1] K. Shojaei and M.M. Arefi, "On the neuro-adaptive feedback linearising control of underactuated autonomous underwater vehicles in three-dimensional space", *IET Control Theory Appl.* 9 (8), pp. 246-258, 2015.
- [2] L. Moreira and C.G. Soares, "H<sub>2</sub> and H<sub>∞</sub> designs for diving and course control of an autonomous underwater vehicle in presence of waves", *IEEE J. Ocean. Eng.* 33 (2), pp. 69-88, 2008.
- [3] R.B. Wynn, V.A.I. Huvenne, T.P. Le Bas, B.J. Murton, D.P. Connelly, B.J. Bett, H.A. Ruhl, K.J. Morris, J. Peakall, D.R. Parsons, E.J. Sumner, S.E. Darby, R.M. Dorrell, and J.E. Hunt, "Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience", *Marine Geology*, vol. 352, pp. 451-468, 2014.
- [4] M. Eichhorn, C. Ament, M. Jacobi, T. Pfuertzenreuter, D. Karimanzira, K. Bley, M. Boer, and H. Wehde, "Modular AUV System with Integrated Real-Time Water Quality Analysis", *Sensors*, 18(6), pp.18-37, 2018.
- [5] P.E. Hagen, N. Storkersen, K. Vestgard, and P. Kartvedt, "The HUGIN 1000 autonomous underwater vehicle for military applications", *Oceans 2003*, 1141-1145 Vol.2, 2003.
- [6] K. Mondal, T. Banerjee, and A. Panja, "Autonomous Underwater Vehicles: Recent Developments and Future Prospects", *International Journal for Research in Applied Science and Engineering Technology*, 7(11), pp. 215-222, 2019.
- [7] F. Zhang, G. Marani, R.N. Smith, and H.T. Choi, "Future Trends in Marine Robotics [TC Spotlight]", *IEEE Robotics & Automation Magazine*, 22(1), pp. 14-122, 2015.
- [8] B. Jalving, "The NDRE-AUV Flight Control System", *IEEE Journal of Oceanic Engineering*, 19(4), pp. 497-501, 1994.
- [9] T. Prestero, Verification of a six-degree of freedom simulation model for the REMUS autonomous underwater vehicle. In Ph.D. Thesis. Massachusetts Institute of Technology, 2001.
- [10] K. Tanakitkorn, P.A. Wilson, S.R. Turnock, and A.B. Phillips, "Depth control for an over-actuated, hover-capable autonomous underwater

- vehicle with experimental verification”, *Mechatronics*, 41(3), pp. 67–81, 2017.
- [11] E.Y. Hong, H.G. Soon, and M. Chitre, “Depth control of an autonomous underwater vehicle, STARFISH”, *OCEANS’10 IEEE SYDNEY*, pp. 1–6, 2010.
  - [12] J. Xu, M. Wang, and L. Qiao, “Dynamical sliding mode control for the trajectory tracking of underactuated unmanned underwater vehicles”, *Ocean Engineering*, 105, pp. 54–63, 2015.
  - [13] N.H. Tran, T.D. Huynh, T.P. Ton, and T.H. Huynh, “Design of depth control for hybrid AUV”, *Proceedings of The 6th International Conference on Advanced Engineering - Theory and Applications 2019*, Bogotá, Colombia, 2019.
  - [14] L. Lapierre, “Robust diving control of an AUV”, *Ocean Engineering*, 36(1), pp. 92–104, 2009.
  - [15] S. Mahapatra and B. Subudhi, “Design and experimental realization of a backstepping nonlinear  $H_\infty$  control for an autonomous underwater vehicle using a nonlinear matrix inequality approach”, *Transactions of the Institute of Measurement and Control*, 40(11), pp. 3390–3403, 2018.
  - [16] U. Ansari and A.H. Bajodah, “Robust generalized dynamic inversion control of autonomous underwater vehicles”, *IFAC Pap. OnLine* 50–1, pp. 10658–10665, 2017.
  - [17] L. Qiao, S. Ruan, G. Zhang, and W. Zhang, “Robust  $H_2$  optimal depth control of an autonomous underwater vehicle with output disturbances and time delay”, *Ocean Engineering*, 165, pp. 399–409, 2018.
  - [18] J.H. Li and P.M. Lee, “Design of an adaptive nonlinear controller for depth control of an autonomous underwater vehicle”, *Ocean Eng.* 32, pp. 2165–2181, 2005.
  - [19] B.W. Yi, L. Qiao, and W.D. Zhang, “Two-time scale path-following of under-actuated marine surface vessels: design and stability analysis using singular perturbation methods”, *Ocean Eng.* 124, pp. 287–297, 2016.
  - [20] E.Y. Hong and M. Chitre, “Roll Control of an Autonomous Underwater Vehicle Using an Internal Rolling Mass”, *Field and Service Robotics. Springer Tracts in Advanced Robotics*, vol 105. Springer, Cham, 2015.
  - [21] T.I. Fossen, *Guidance and Control of Ocean Vehicles*. JohnWiley and Sons, 1995.



# Implementation and Enhancement of Set-Based Guidance by Velocity Obstacle along with LiDAR for Unmanned Surface Vehicles

Ngoc-Huy Tran✉

*Ho Chi Minh City University of  
Technology, VNUHCM*  
Ho Chi Minh City, Vietnam  
tnhuy@hcmut.edu.vn

Minh-Hung Vu

*Petro Vietnam University, PVN*  
Ho Chi Minh City, Vietnam  
hungvm@pvu.edu.vn

Tu-Cuong Nguyen

*Ho Chi Minh City University of  
Technology, VNUHCM*  
Ho Chi Minh City, Vietnam  
tucuongbrt@gmail.com

Minh-Tam Phan

*Ho Chi Minh City University of Technology,  
VNUHCM*  
Ho Chi Minh City, Vietnam  
1613063@hcmut.edu.vn

Quang-Ha Pham

*Ho Chi Minh City University of Technology,  
VNUHCM*  
Ho Chi Minh City, Vietnam  
1610864@hcmut.edu.vn

**Abstract**— In hazardous conditions, the use of an autonomous unmanned surface vessel (USV) may be a solution to an obstacle avoidance problem with minimal human intervention but maximal efficiency. This article describes a method of utilizing Velocity Obstacle algorithm to enhance collision-free paths of Set-Based Guidance (SBG) for USV equipped with Light Detection and Ranging (LiDAR) for detecting obstacles. Additionally, we also propose a novel trapezium path with three waypoints instead of a single waypoint as conventional SBG. Moreover, we build a hardware architecture for USV to realize the obstacle avoidance algorithm in practice. To verify safety, efficacy, and pragmatics, simulations are conducted in situations of static obstacles with different shapes and moving ship in three typical collision scenarios: head-on, overtaking, and crossing. Finally, experiments with a real static are carried out and experimental results prove that the advanced SBG is able to guide USV to avoid collision automatically.

**Keywords**— *Unmanned Surface Vehicle, Obstacle Detection, Obstacle Avoidance, Velocity Obstacle, Set-Based Guidance, LiDAR.*

## I. INTRODUCTION

Today, unmanned surface vehicle (USV) attracts wide attention due to its flexibility, portability, stability and increasingly high intelligence. It is used as a means for environmental monitoring, shallow water surveying, military reconnaissance, homeland security and has become a hot research topic for intelligent ships around the world [1]. Typically, with a path-following task, the guidance system consists of laws for heading, surge velocity to ensure convergence to the desired path. Then the control system calculates the thruster force to track the reference states by the guidance system. A common approach for path following is the Line of Sight (LOS) method, which allows USV to follow straight paths [2], [3] and curved paths [4]-[5]. However, in some situations, there are many unknown obstacles which hinders path-following algorithms to be carried out successfully.

Collision avoidance is an important research aspect in the development of USV, with the aim to plan a collision-free path in unknown environments [6]. The local collision-free path can be generated by two approaches, including path

searching-based local path planning method, which plans the paths or target waypoints tracked by USV, and the behaviour-based reactive obstacle avoidance method, which generates guidance velocity and guidance angle based on obstacle deeds [7]. Kuwata [8] has proposed a behaviour-based reactive approach, which adopt Velocity Obstacle (VO) to define a cone space of velocity. Kim [9] has implemented genetic algorithm to optimize collision-free path with three objective functions of travel time, environmental loads, and avoiding obstacles. Zhao [10] has introduced a novel approach of real-time automatic avoiding collisions of multiple vessels in compliance with COLREGs rules. Moe and Pettersen [11] have proposed set-based method with path following task and obstacle avoidance task. Myre [12] has modified set-based method to be more appropriate for underactuated systems like USV by using an additional obstacle avoidance waypoint. In this paper, we propose a more effective way of setting new waypoints, which uses three waypoints corresponding three vertices of a trapezium instead of single one. These waypoints are calculated online at the time as avoiding obstacle by adopting velocity algorithm (VO), which allow USV avoiding static and dynamic obstacles.

This paper is organized as follows: The suggested set-based guidance is presented for collision avoidance in Section 2. Finally, the results are given in Section 3 and conclusions in Section 4..

## II. ADVANCED SET-BASED GUIDANCE

### A. Concept of Velocity Obstacle

Velocity Obstacle (VO) approach for generating a collision-free path has been proposed by Kuwata [8]. The basis idea of this algorithm is to turn a general single obstacle avoidance problem to a static one. To simplify the problem, a shape of the obstacle is temporarily ignored and replaced by its centroid and a safety circle with radius of  $R_s$  (Fig. 1). It can be inferred from Fig.1 that USV will collide with the target ship when belongs to the cone space  $T-UT^+$  which is formed by two tangents of the safety circle. As a result, the collision between USV and the target ship will not take place when a following condition is satisfied:

$$\begin{aligned} \arg(\vec{v}_{uo}) &\notin \text{ConeSpace}(\vec{R}_{uo}, R_s) \\ &= (\arg(\vec{R}_{uo}) - \delta; \arg(\vec{R}_{uo}) + \delta) \end{aligned} \quad (1)$$

where

$\arg(\cdot)$ : is an argument of vector ( $\cdot$ )

$\delta = \text{asin}(R_s / \sigma)$ : is a half of cone space angle.

$R_s$ : is a radius of the safety circle.

$\vec{R}_{uo} = O - U$ : is a vector from the obstacle (O) to USV (U)

$\sigma = \|\vec{R}_{uo}\|$ : is distance between USV and the obstacle.

$\|\cdot\|$  operator is Euclidean norm of ( $\cdot$ ).

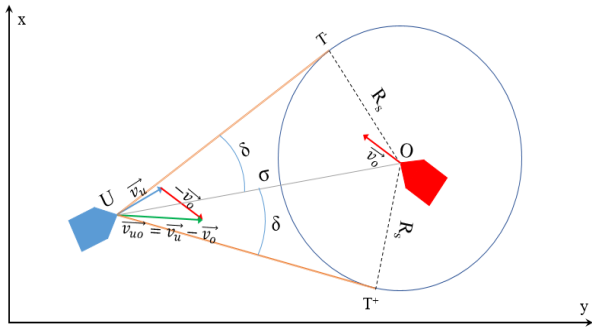


Fig. 1. Illustration of parameters used for assessing and avoiding collision according to Velocity Obstacle.

Once it is anticipated for a collision between USV and the obstacle according to (1), USV has to change its course to a desired heading  $\psi_{ud}$  such that a new relative velocity  $\vec{v}_{ud}$  is out of the cone space. Nevertheless, changing the course angle will cost USV amount of energy as much as its inertia, for the aim of minimizing course change,  $\arg(\vec{v}_{ud})$  is equal to  $\vec{R}_{UT^-}$  or  $\vec{R}_{UT^+}$  respect to collision scenarios that are regulated by COLREGs rules.

Now, we have:

$$\begin{aligned} \begin{cases} \|\vec{v}_u\| \cos(\psi_{ud}) - \|\vec{v}_o\| \cos(\psi_o) = \|\vec{v}_{ud}\| \cos(\arg(\vec{v}_{ud})) \\ \|\vec{v}_u\| \sin(\psi_{ud}) - \|\vec{v}_o\| \sin(\psi_o) = \|\vec{v}_{ud}\| \sin(\arg(\vec{v}_{ud})) \end{cases} \\ \Rightarrow \psi_{ud} = \text{asin}\left(\frac{v_o}{v_u} \sin(\psi_o - \arg(\vec{v}_{ud}))\right) + \arg(\vec{v}_{ud}) \end{aligned} \quad (2)$$

However, the desired yaw  $\psi_{ud}$  is not sufficient to avoid the obstacle because of uncertainty in measuring the obstacle pose. Hence, the uncertainty should be concerned in calculating. Let measurement vector of the obstacle as follow:

$$[\hat{x}_o, \hat{y}_o, \hat{v}_o, \hat{\psi}_o]^T = [x_o + \delta_x, y_o + \delta_y, v_o + \delta_v, \psi_o + \delta_\psi]^T \quad (3)$$

where  $[x_o \ y_o \ v_o \ \psi_o]^T$  is the real state of the obstacle,  $[\delta_x \ \delta_y \ \delta_v \ \delta_\psi]^T$  is a vector of the uncertainty and  $|\delta_x|, |\delta_y| \leq r_p, |\delta_v| \leq r_v < v_o, |\delta_\psi| \leq r_\psi$ .

It can be seen that the real position of the obstacle  $(x_o, y_o)$  lies in a circle which has radius of  $r_p \sqrt{2}$  and a center  $(\hat{x}_o, \hat{y}_o)$ . To ensure safety navigation, the safety radius would better to be extended to new radius of  $(R_s + r_p \sqrt{2})$ . Next, the uncertainty of  $\delta_v$  and  $\delta_\psi$  is handled by find a velocity  $v_o^{wc}$  and a heading angle  $\psi_o^{wc}$  that require USV to make the largest change in its course angle in the worst case. Partially derivating (2) respect to  $v_o$  and  $\psi_o$ , we have:

$$\begin{cases} \frac{\partial \psi_{ud}}{\partial v_o} = \frac{\sin(\psi_o - \arg(\vec{v}_{ud}))}{\sqrt{v_u^2 - v_o^2 \sin^2(\psi_o - \arg(\vec{v}_{ud}))}} \\ \frac{\partial \psi_{ud}}{\partial \psi_o} = \frac{v_o \cos(\psi_o - \arg(\vec{v}_{ud}))}{\sqrt{v_u^2 - v_o^2 \sin^2(\psi_o - \arg(\vec{v}_{ud}))}} \end{cases} \quad (4)$$

Obviously, Eq. (4) is impossible to have real solutions, so a solution that USV make the largest course change locates on edges of problem domain, i.e:

$$(v_o^{wc}; \psi_o^{wc}) \in \{(\hat{v}_o \pm r_v; \hat{\psi}_o \pm r_\psi)\} \quad (5)$$

### B. Obstacle avoidance waypoints

In an ideal condition, when the heading angle can reach immediately a value of the desired heading angle  $\psi_{ud}$  and the surge velocity is still unchanged, USV is able to avoid obstacle according to the concept of VO algorithm and obstacle avoidance path is a straight line with a slope of  $\tan(\psi_{ud})$ . Apparently, this ideal condition will never be satisfied by a practical USV model with motion constraints, but USV has capacity of avoiding obstacle when following the ideal obstacle avoidance path.

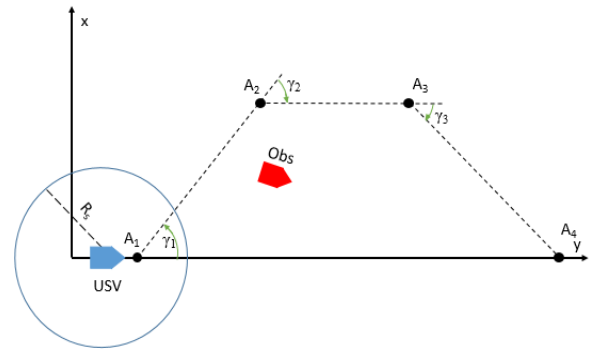


Fig. 2. The trapezium path and its parameters.

A complete obstacle avoidance algorithm usually includes two tasks of avoiding collision and path following. Along with avoiding hazardous, USV has to keep distance as close as possible with the global path to facilitate the path following task. In this paper, we proposed a trapezium collision-free path (Fig. 2) with parameters determined as follow:

$$\begin{cases} A_i = U(t_i), i = 1 \div 3 \\ \gamma_1 = \psi_{ud}(t_1) \\ \gamma_2 = \alpha_k \\ \gamma_3 = \alpha_k - \text{sign}(\gamma_1 - \alpha_k) \mu \end{cases} \quad (6)$$

where  $\alpha_k$  is a slope angle of the global path,  $\psi_{ud}$  is the desired heading angle calculated according to VO,  $\mu = \pi/4$  for

head-on and  $\mu = \pi/6$  for overtaking and crossing,  $t_1$  is time to turn path following task into obstacle avoidance task when  $\sigma$  is less than a given threshold,  $t_2$ ,  $t_3$  is the moment the relative velocity  $\bar{v}_{uo}$  satisfying (1) but with a assumption that the heading angle  $\psi_u$  is equal to  $\gamma_2$ ,  $\gamma_3$ , respectively.

### III. RESULTS AND DISCUSSION

In the previous sections, we presented the dynamic model of USV, identifying obstacles using LiDAR, obstacle avoidance algorithm and hardware structure to realize the algorithm. To test the algorithm's ability to navigate safely, several simulations were performed with static obstacles collision situation with three different shapes and three collision situations with dynamic obstacles including: head-on, overtaking and crossing. Then, an experiment is carried out to evaluate the obstacle avoidance algorithm with a real static obstacle.

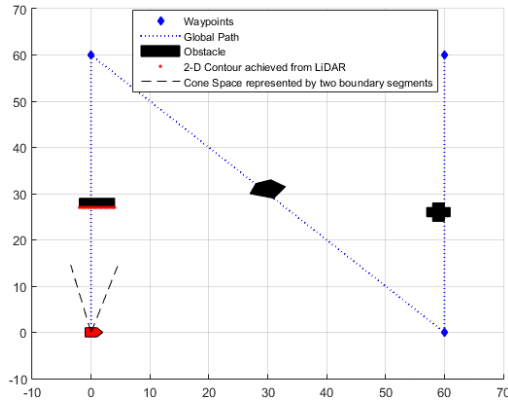
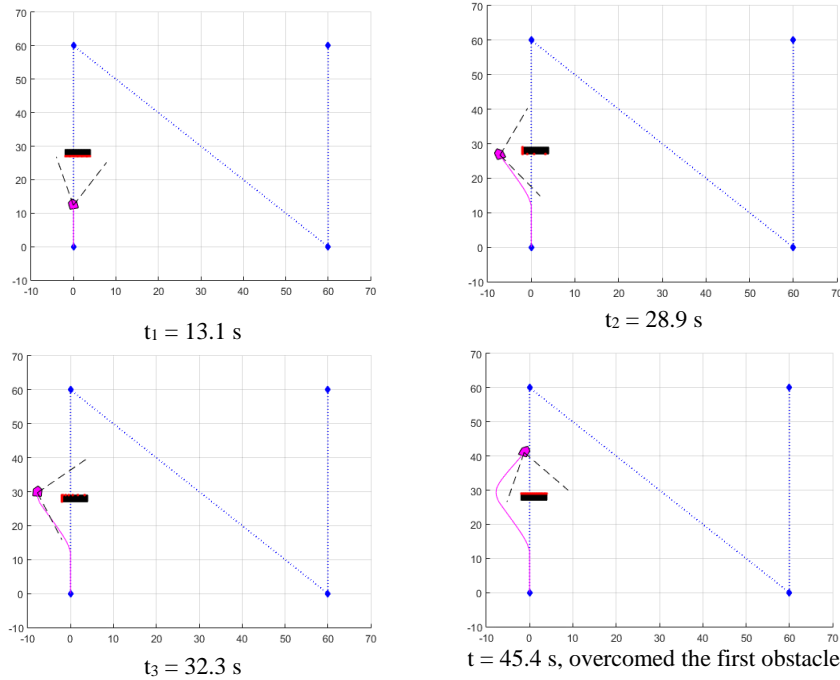


Fig. 3. Notations for the algorithm



#### A. Avoiding static obstacles simulation

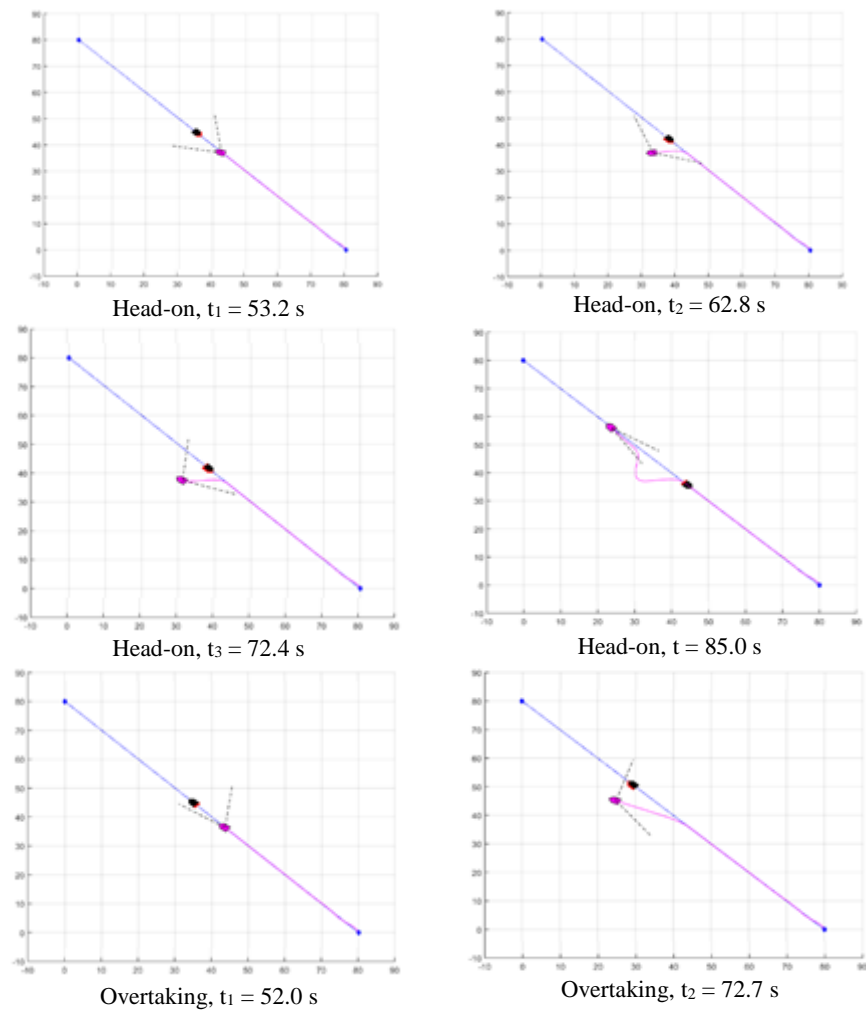
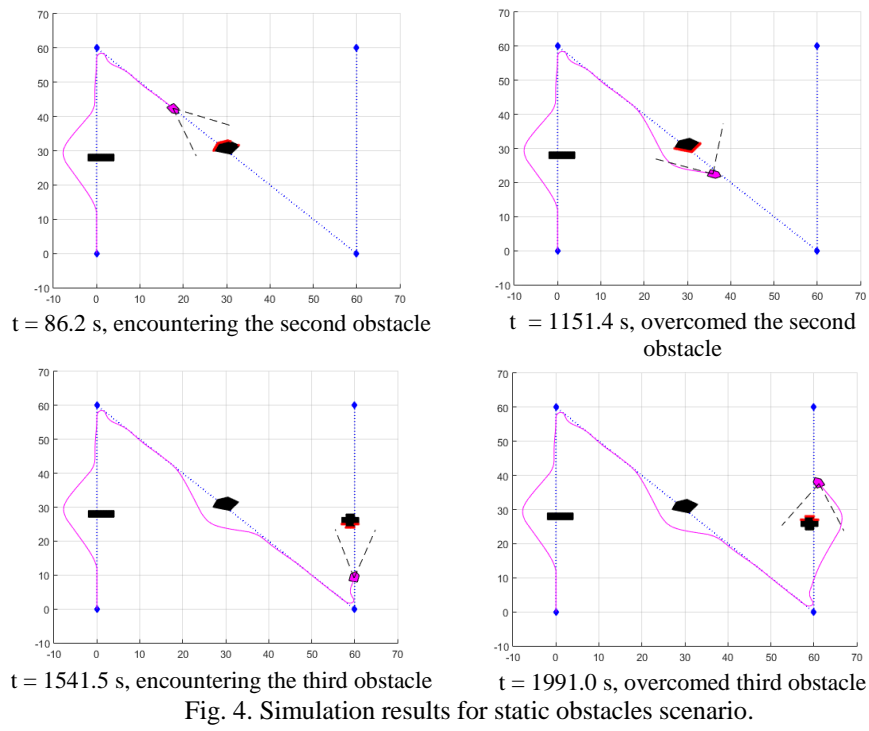
In this simulation, USV is located at position (0,0), heading angle is  $\pi/2$  and begins to follow the global path (blue dot) created by waypoints (blue diamond) as Fig 3. Three black static obstacles with the shape of a rectangle, pentagon and cross are preset on the global path. From the profile of the scanning obstacle used LiDAR, a cone space is defined by VO and is represented by two boundary dash segments.

The results in Fig. 4 pinpoint out that this algorithm will be effective and pragmatic when it is implemented to avoid obstacles. Convergence to the global path after detecting and avoiding obstacles reassures that the obstacle avoidance algorithm did not affect control and following path of USV. Moreover, the simulations verify safe navigation of the algorithm when USV keeps distance to obstacle over time of overcoming and optimization in choosing side to pass such that the course change is as minor as possible.

#### B. Avoiding dynamic obstacles simulations

Conventionally, the moving obstacles that USVs encounter at sea are ships or others surface vehicles, so obstacle avoidance must comply with the COLREGs rules. However, safety and optimization of evasive manoeuvres are maintained. Simulation results in Fig. 5 show that USV returns promptly to the global path after overcoming the obstacles. Hence, the avoiding obstacle process does not affect significantly to control and others tasks.

Throughout the results and analyses, we arrive at a conclusion that proposed algorithm is capable of dealing with both static obstacles and dynamic obstacles. It guides USV to avoid obstacle safely with short path length and does not hinders others missions.



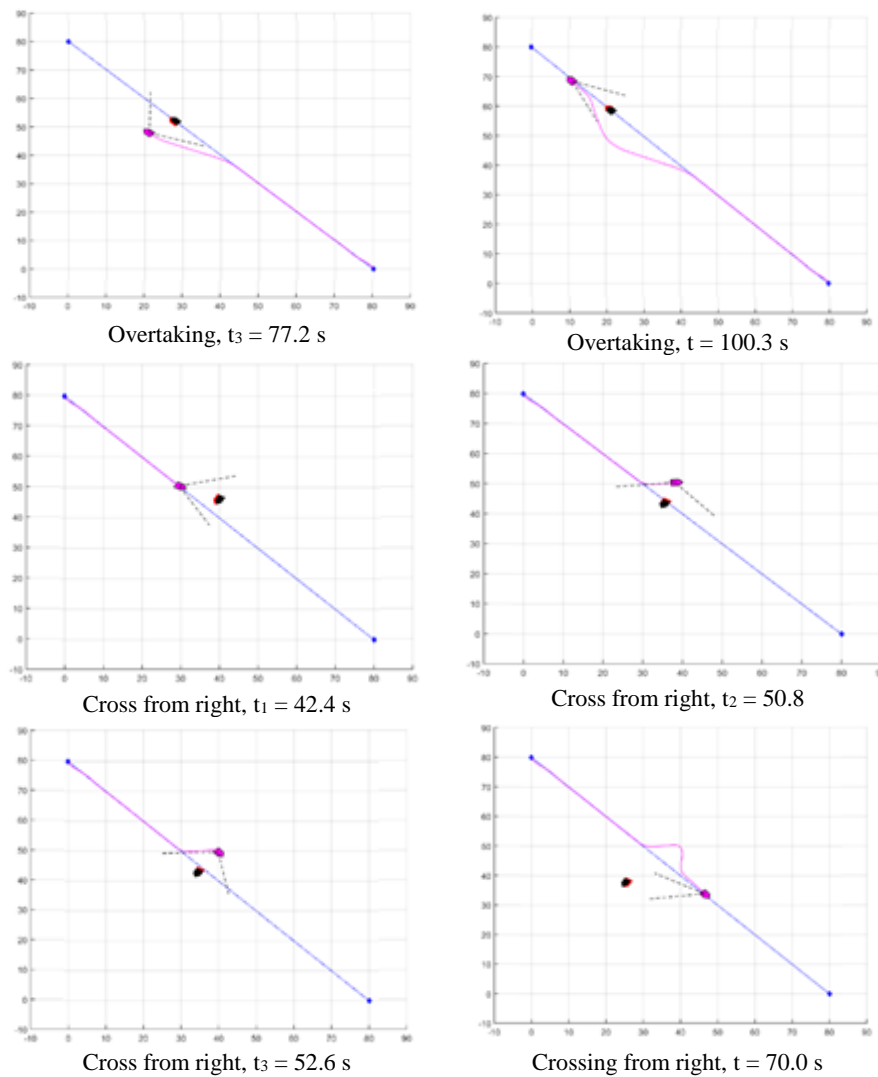


Fig. 5. Dynamic obstacles simulation results for three scenarios: head-on, overtaking, and cross from right

#### IV. CONCLUSION

This paper has reviewed on some studies about obstacle avoidance including the path searching-based method and behaviour-based reactive method. Obstacle detection with high accuracy and reliability using LiDAR has been introduced to improve quality of the obstacle avoidance task. We have also proposed the advanced set-based guidance with the trapezium collision-free path that is planned by adopting the concept of Velocity Obstacle and LiDAR measurement results. The simulation results have confirmed that this algorithm is safe, effective and pragmatic to both static and dynamic obstacles. Moreover, we have adopted successfully this theory in avoiding real static obstacles.

In this study, we only concentrate on studying the obstacle avoidance algorithm for a single obstacle while multiple obstacles at the same should be concerned in practice. Moreover, with proposed hardware architecture of USV, we are looking forward to realizing the obstacle avoidance algorithm with dynamic obstacles.

#### ACKNOWLEDGMENT

This research is funded by Petro Vietnam University, PVN, under grant number GV2005 and supported by

Laboratory of Advance Design and Manufacturing Processes, HCMUT.

#### REFERENCES

- [1] Campbell, S.; Naeem, W.; Irwin, G.W. "A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres". *Annu. Rev. Control* 2012, 36, 267–283.
- [2] T. I. Fossen, M. Breivik, and R. Skjetne, "Line-of-sight path following of underactuated marine craft," *Proc. 6th IFAC Conference on Manoeuvring and Control of Marine Craft*, pp. 244–249, 2003.
- [3] S. Oh and J. Sun, "Path following of underactuated marine surface vessels using line-of-sight based model predictive control" *Ocean Engineering*, 2010, vol. 37, no. 2-3, pp. 289–295.
- [4] R. Skjetne, U. Jorgensen, and A. R. Teel, "Line-of-sight path-following along regularly parametrized curves solved as a generic maneuvering problem," in *Proc. IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 2467–2474.
- [5] E. Borhaug and K. Y. Pettersen, "LOS path following for underactuated underwater vehicle," in *Proc. 7th IFAC Conference on Manoeuvring and Control of Marine Craft*, 2006.
- [6] Y. Liu and R. Bucknall, "Path planning algorithm for unmanned surface vehicle formations in a practical maritime environment," *Ocean Eng.*, vol. 97, pp. 126–144, Mar. 2015.
- [7] Wang, Y., Yu, X., Liang, X., & Li, B. "A COLREGs-based obstacle avoidance approach for unmanned surface vehicles." *Ocean Engineering*, 2018, pp. 110–124.



- [8] Kuwata, Y., Wolf, M.T., Zargitsky, D., Huntsberger, T.L., "Safe maritime navigation with COLREGS using velocity obstacles." In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. IEEE, pp. 4728–4734.
- [9] Kim, H., Kim, S.H., Jeon, M., Kim, J.H., Song, S., Paik, K.J., "A study on path optimization method of an unmanned surface vehicle under environmental loads using genetic algorithm.", Ocean Eng. 2018, , 616–624.
- [10] Y. Zhao, W. Li, P. Shi, "A real-time collision avoidance learning system for Unmanned Surface Vessels", Neurocomputing 182 (2016) 255–266.
- [11] S. Moe, K.Y. Pettersen, "Set-based line-of-sight (LOS) path following with collision avoidance for underactuated unmanned surface vessel", 24th Mediterr. Conf. Control Autom. MED 2016, Athens, Greece, 2016, pp. 402–409.
- [12] Myre, H., 2016. Collision Avoidance for Autonomous Surface Vehicles Using Velocity Obstacle and Set-Based Guidance (Master's thesis). NTNU.

# Design Integrated Staff Welcoming and Administration System Based on Facial Recognition for Smart University

Dat Tan La

*Faculty of Advanced Science and Technology*  
*Danang University of Science and Technology*  
Danang, VietNam  
ltdd117@gmail.com

Huy Quang Tran

*Faculty of Advanced Science and Technology*  
*Danang University of Science and Technology*  
Danang, VietNam  
quanghuyphiet@gmail.com

Nhat Tien Le

*Faculty of Advanced Science and Technology*  
*Danang University of Science and Technology*  
Danang, VietNam  
ltn281097@gmail.com

Quang Luong Nguyen

*Faculty of Advanced Science and Technology*  
*Danang University of Science and Technology*  
Danang, VietNam  
quang.cl.qh@gmail.com

Thu Thi Anh Nguyen

*Faculty of Advanced Science and Technology*  
*Danang University of Science and Technology*  
Danang, VietNam  
ntathu@dut.udn.vn

Tuan Van Pham

*Faculty of Advanced Science and Technology*  
*Danang University of Science and Technology*  
Danang, VietNam  
pvtuan@dut.udn.vn

**Abstract**—In recent years, facial recognition technology is not only applied for security, healthcare, business but it is also exploited creatively in education and for smart university development. In this study, an Automated Employee Attendance Management System has been designed and implemented in a university's building for welcoming staff and supporting university administration. The system starts with a facial detection module which is based on the pre-trained Multi-Task Cascaded Convolutional Network model. Then the feature vectors are created by using the ResNet34 network which results in the 128-dimension embedding vector. Face recognition is carried out by using various techniques such as K-Nearest Neighbors, Support Vector Machine. Besides the welcoming front-end module, a web-based application for querying information to manage people entering the building is also built as a back-end module. The proposed face recognition models have been trained and tested on a collected face database and a self-built face database of employees who work at the university building. The evaluation results show high recognition rates in terms of Precision, Recall, Accuracy, F1-score and reasonable processing time. The proposed system has been piloted at the university for further development and research on face recognition technologies and smart building management.

**Keywords**—Facial recognition, Multi-task Cascaded Convolutional Neural Network, Resnet34, Support Vector Machine, Automated Attendance Management System.

## I. OVERVIEW

### A. Related works

When building a digitally enabled working environment, a sophisticated platform for attendance management becomes necessary. As a center of learning and excellence, higher education institutions cannot afford to be left behind in using and developing these new technologies. One of many approaches to develop non-intrusive automated employee attendance management systems based on human facial recognition. This technology has been identified as a great motivation and also a challenge for society, policymakers,

smart governments and cities as stated in an AI Now Academy report [1]. Many challenges that face recognition technologies may face, such as high accuracy, short processing times for a real-time application, ability to identify in unfavorable conditions such as partially obscured, expressions, viewing angles, lighting conditions, etc.

Many methods have been developed to help detect and identify faces. Currently, the two most widely used methods for detecting human faces in images are the Haar Cascade [2] and HOG characteristics [3]. Besides, according to another study in [4], the combination of Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) has been proved to be highly effective. According to the article [5], the Haar Cascade method has shown that the processing speed is better than that of HOG in face detection. The detection of objects using cascade based on Haar Cascade features is an effective and fast method of object detection using simple features Boosted Cascade [6], [7]. In recent years, the Convolutional Neural Network (CNN) has been shown to potentially outperform all classical approaches based on standard features due to its generalization ability [8]. Recently, one of the most widely-used was the Multi-Task Cascaded Convolutional Network (MTCNN) detector [9], [10], which performs both face detection and face landmarking. So far, the MTCNN is still proposed to be used in the state-of-the-art face recognition system. In addition to this model, the ResNet34 neural network [11] has been developed to effectively serve to extract facial features in form of 128-dimensional vectors that are specific to all facial features. These feature vectors are then labeled and serve to match and predict the input image [12].

### B. Automated Employee Attendance Management System

A proposed automated employee attendance management system (AEAMS) consists of multiple modules which is shown in the Fig. 1. Firstly, the face image will go through the face recognition module. Then the ID output will be matched to existing datasets including ID lists and timetables. Finally,

welcoming information will be extracted and displayed via the screen. The system also stores the image and the entrance time.

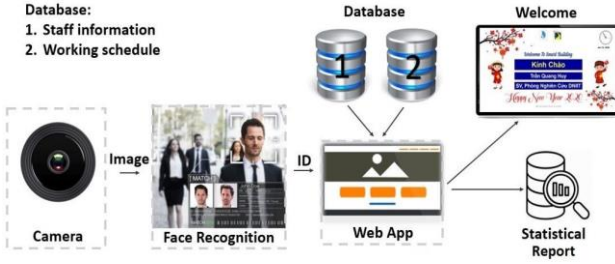


Fig. 1. Block diagram of the AEAMS.

Details of the techniques used are described in Section 2. The used data includes the online dataset and the self-built dataset are presented in Section 3. Test results and discussions will be presented in Section 4. Finally, Section 5 is the conclusion.

## II. PROPOSAL APPROACHES

The face recognition module has been developed as shown in Fig. 2. The input face image is detected by the MTCNN model. Feature extraction will be carried out by the ResNet34 model to extract the prominent features of the face. At the training stage, three various models named as Euclidean distance measure, K-Nearest Neighbors, Support Vector Machine have been trained. During the testing phase, decoding process will rely on the characteristic vectors that have been previously trained to make predictions.

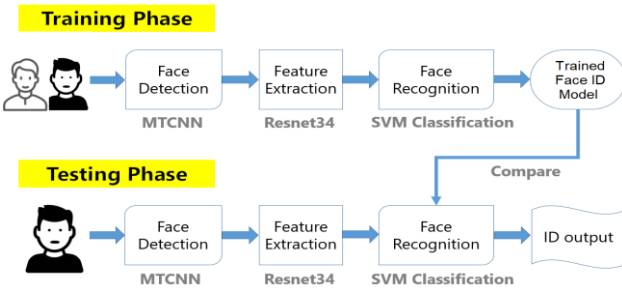


Fig. 2. Block diagram of the recognition process.

### A. Face Detection

The MTCNN model is called a multi-task network because each of the three models in the cascade (P-Net, R-Net and O-Net) are trained on three tasks, e.g. make three types of predictions; they are: face classification, bounding box regression, and facial landmark localization. The three models are not connected directly, instead, outputs of the previous stage are fed as input to the next stage. This allows additional processing to be performed between stages. For example, non-maximum suppression is used to filter the candidate bounding boxes proposed by the first-stage P-Net prior to providing them to the second stage R-Net model. The MTCNN architecture is reasonably complex to implement.

### B. Feature Extraction

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note

peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

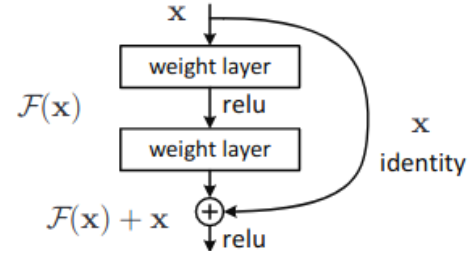


Fig. 3. Structure of a building block [11]

With the Residual block, it is possible to train CNN models with "bigger" dimensions and complexity without worrying about exploding / vanishing gradients. The key to the Residual block is that after every 2 layers, we add input to the output:  $F(x) + x$ . ResNet is a CNN network composed of many small Residual blocks formed. Formally, denoting the desired underlying mapping as  $H(x)$ , we let the stacked nonlinear layers fit another mapping of  $F(x)$ :

$$F(x) = H(x) - x \quad (1)$$

The original mapping is recast into  $F(x) + x$ . According to [11], [13] Resnet network model with  $x$  (shortcut) shortcut has proved the effectiveness compared to previous models such as VGG16 [14] in terms of training effectiveness. When the number of classes in the model increased with the error in the training (top-1 error) reduced by 5%, the top-5 error was only around 6.71%. In [15] recognition performance was also increased to 5 % with much lower memory usage than VGG16.

### C. Recognition Models

#### 1) Centroid-based Euclidean distance measurement:

To measure similarity between two feature vectors, Euclidean distance measurement has been applied as follows:

$$D_e = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2)$$

Where:

$n$  = number of dimensions

$p_i, q_i$  = data points

To find the centroid point representing feature vectors of one face ID, the first feature vector of a class will be assigned as the centroid feature vector of the class. From the second feature vector and over, the centroid feature vector is the average value of the feature vector and the current value of centroid feature vector. In the end, there will be only an average centroid feature vector for each class in the training set.

One big advantage of this method is the very fast computational time and fast training time. In contrast, the performance of the model using this method is not as good as other methods and the model is sensitive to noise. Another constraint is to classify a person not in the training set, we

must use a threshold. This threshold can only be chosen by experimenting. It changes according to the data set.

### 2) K-Nearest Neighbors algorithm:

The principle behind nearest neighbor methods is to find a pre-defined number of training samples closest in distance to the new point, and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning). The Euclidean distance measure is the most common choice. Neighbors-based methods are known as non-generalizing machine learning methods, since they simply “remember” all of its training data [16].

To train a good KNN model, we must choose a correct K number of neighbors, which is based on the training set. The KNN configuration includes KDTree learning algorithm, K number of neighbors, Euclidean distance metric, uniform weight metric which means different distances between neighbors have the same weight. The algorithm will find a query point for each data class. After the training, we have a model which is the representative of the training set. The training process of KNN models is very fast because the model only remembers the data points, not actually doing many calculations. Because of that, it takes much time for the model to predict a feature vector.

### 3) Support Vector Machine:

The SVM is used because it works well on data with high dimension [17]. The algorithm will find a hyperplane in a 128-dimensional space that distinctly classifies the data points. The hyperplane is chosen under constraint of maximizing the margin between data points and the hyperplane. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, the margin of the classifier is maximized. After training and testing on many SVM model configurations which is a combination of Linear or Radial Bias Function kernel and Regularization Parameter C are 0.01, 1 and 100, the result show that, with RBF kernel and regularization parameter C = 100, the model gives the best performance in all the configurations.

### D. Post-Processing

A post-processing module has been developed to support decision making based on recognition's result of a sequence of faces, not a single face. In case that there is more than one individual in an image frame, grouping of the faces belonging to the same person is essential to be done prior to the post-processing.

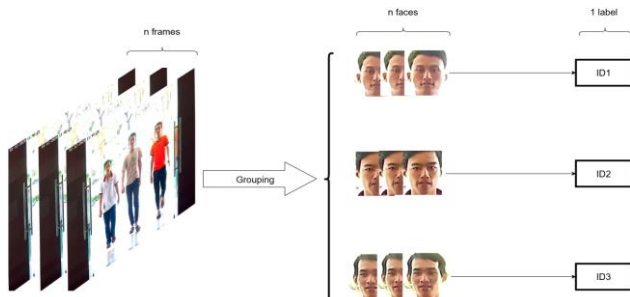


Fig. 4. The process of post-processing.

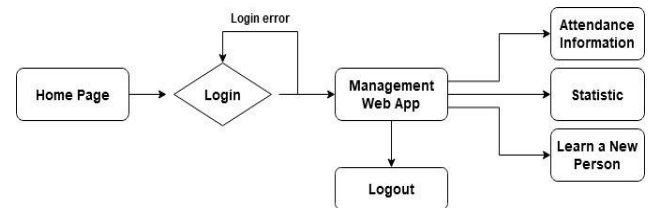
To do this post-processing, we use the extracted feature vector as a measurement of similarity for each face in the

frame. The same person must have features vectors closest to each other. The Euclidean distance measure has been exploited for similarity measurement as depicted in Fig. 4.

Firstly, detection of faces in the frame is carried out followed by their feature extraction for each face. Secondly, each detected face in the first frame is assigned to a group. Next, for each recurrent frame, the above process is repeated. For each face in the frame, we measure the similarity between the feature vector of this face and the features vectors of groups. We assign a face to the group which is the most similar (minimum Euclidean distance) to the face. Finally, each group has a live time and a position which is represented by the last face of that group. If the live time or the position exceed predefined values, the group is removed. In the project, we define live time as 3 seconds which is the average time for a person to cross the recognition zone.

### E. Web-based Management

A web application has been developed and integrated with the system to display information of the detected face about names, positions and units. For the case of many people entering the building gate at the same time, the screen will greet each person alternatively. In addition to the welcoming webapp, a staff management webapp has been developed as shown in Fig. 5. This application provides the following information: Security features login, logout; Attendance information (including name, unit, position, image) of an individual entering the building in chronological order and tracking images at the time that person enters the building; Statistics of staff frequency working in the building by the



chart; New enrollment to the system.

Fig. 5. Functional Structure Diagram and Management Interface.

## III. DATABASES AND TRAINING

### A. Facial online dataset

The FERET dataset [18] includes 14,126 images of 1199 individuals collected from 1993 to 1996. This database was designed for supporting train and test in large numbers of appearance changes of the individuals. Each person has about 10 images, 7 images are used for the training set and 3 images are used for the testing set.

### B. Self-built DUT dataset

#### 1) Collect video data:

This self-built dataset includes photos of the faces of employees working at the university building. This data is collected naturally with situations that come from all directions of moving employees during the working time. Employees can go to the entrance from different directions such as A, B, C, D, E, and then move ahead into the building via 3 different directions as noted with arrows 1, 2, 3. (Fig. 6.)

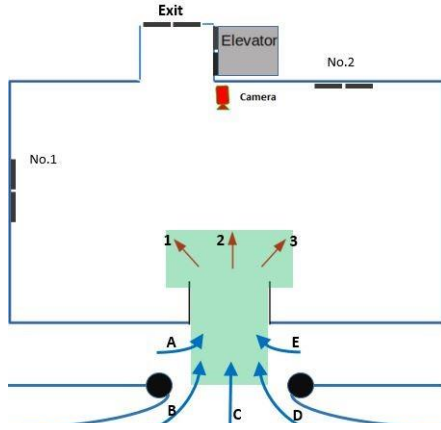


Fig. 6. Data collection scenario.

Face images will be cut out from frames of videos by module face detection. Then, the dataset is divided into three subsets: a training set, a test set in well-matched (WM) condition and a test set in highly-mismatched (HM) condition.

- Training set: It includes front face images and face slightly tilted to the left or right, image quality, and bright conditions are relatively good.
- WM set: Their contents and recording conditions are very similar to the training set. Sometimes the face shows a slight smile.
- HM set: There are many images of face blurred, face deviation angle from the main direction large (30 - 45), head down, eyes closed, hard lighting condition (too bright), smiling expression on face.

The original dataset was built from 80 individuals, each person has at least 10 photos. Description of the DUT dataset is shown in TABLE I.

TABLE I. DESCRIPTION OF THE DUT DATASET

Number of people	Train	Test		Sum
		WM	HM	
80	1561	661	9466	11688

## 2) Data Augmentation:

Data augmentation technique is used to increase data to about 1,000 images for each individual in the DUT dataset. The following operations are applied to expand the original dataset:

- Skew or tilt an image either left, right, forwards, backwards or by a random corner. The image will be skewed by a random amount in different directions.
- Elastic Distortions can make distortions to an image while maintaining the image's aspect ratio. It will be only used with a small margin to prevent the face from changing too much.
- Random Erasing is a technique used to make models robust to occlusion. We realized that some of the images after the face detection phase have the face partially obscured.

- Other operations are performed such as: changing the brightness, contrast of the image, rotating without crop, elastic distortion.

After Data Augmentation, we have a data set of 104000 images of 80 different individuals working in the building. The expanded data is shown in TABLE II.

TABLE II. DESCRIPTION OF THE EXPANDED DATASET

Number of people	Train	Test		Sum
		WM	HM	
80	56000	24000	24000	104000

## IV. SYSTEM SETUP ANG IMPLEMENTATION

### A. Integrated System

In order to meet the requirements of the university as the user unit, the team designed and implemented an integrated system for managing people in and out of the building based on facial recognition technology as the diagram in Figure 7.

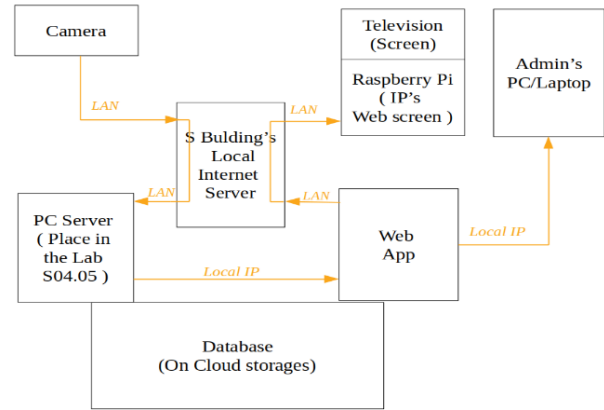


Fig. 7. Block diagram connecting devices in hardware systems

The system includes video recording, data transmission, automated face recognition software, information display software on the welcome interface, and management software that provides monitoring information for authorized usage.

The managed web application provides the following information: Security features login, logout; All information (including name, unit, position, image) of an individual entering the building in chronological order and tracking images at the time that person enters the building; Statistics of staff frequency working in the building; Features for enrollment new staff into the AEAMS.

## V. EXPERIMENTAL RESULTS AND EVALUATION

### A. Performance Criteria

In order to evaluate recognition performance, the following metrics are used:

$$Recall(RE) = \frac{TP}{TP + FN} \quad (3)$$

$$Precision(PRE) = \frac{TP}{TP + FP} \quad (4)$$

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$



$$F1 - Score(F1) = \frac{2 * PRE * RE}{PRE + RE} \quad (6)$$

where TP: true positives; FP: false positives; FN: false negatives; TN: true negative.

### B. Result for Face Recognition

#### 1) Test 1:

The first recognition model based on MTCNN, ResNet34 was tested on the FERET dataset. Output features extracted from the ResNet34 were used directly for classification using Euclidean distance measure. We varied the number of people of the dataset to assess its influence on the model performance. As shown in TABLE III, accuracy was reduced by about 4% with the number of people being 1000, compared to testing on 50 people. As can be seen, the face recognition model has also achieved relatively good recognition performance of about 90%.

TABLE III. EVALUATION RESULTS ON THE FERET

Model	Criteria	Number of people		
		50	500	1000
MTCNN + ResNet34 + Euclidean	RE	94,63%	92,11%	90,60%
	PRE	96,82%	94,90%	94,24%
	ACC	94,63%	92,11%	90,60%
	F1	95,00%	92,30%	91,05%

#### 2) Test 2:

Three more recognition models, as depicted in TABLE IV, have been evaluated through two different testing scenarios named as WM and HM. The derived results show that, on the WM testing set, three algorithms all give high evaluation scores of about or higher 90%. Moreover, the KNN and SVM models give an accuracy gain of 9-10% in comparison to the model that uses Euclidean distance measure. This proves that KNN and SVM model is more reliable while the Euclidean distance model has an acceptable accuracy. The accuracy error rate of about 1.5% with the SVM model mainly comes from only 2 individuals that the model can't distinguish between them.

On the HM testing set, there is a reduction of about 22% in recognition performance by the Euclidean distance model. However, the KNN and SVM model maintain quite good performance on this HM testing set. The accuracy difference obtained between KNN and SVM is about 4% which show slightly better recognition performance of the SVM model.

TABLE IV. EVALUATION RESULTS ON THE SELF-BUILT DUT DATASET (%)

Model	Scenario	RE	PRE	ACC	F1
MTCNN + ResNet34 + Euclidean	WM	90.24	89.65	89.65	89.94
	HM	71.91	67.71	67.71	69.74
MTCNN + ResNet34 + KNN	WM	97.59	97.38	97.38	97.48
	HM	80.90	78.85	78.85	79.86
MTCNN + ResNet34 + SVM	WM	98.71	98.55	98.55	98.62
	HM	85.50	82.97	82.97	84.21

#### 3) Test 3:

The processing time of each main module and the total processing time of the whole system are shown in Fig. 8. These results have been obtained by testing the system on the following hardware computer: 4.1 GHz core i5-9400F, GPU mode with 16.00 GB RAM, 6GB VRAM GPU NVIDIA RTX2060.

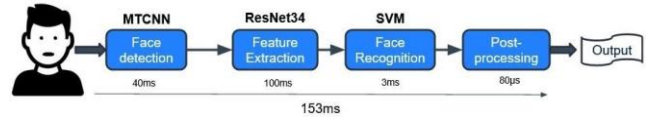


Fig. 8. Computational time.

## VI. CONCLUSION

In this study, an automated employee attendance management system (AEAMS) based on face recognition technology has been developed, installed and evaluated through different testing scenarios. Various algorithms and deep learning models have been designed for building the face recognition module. Because of the requirements of high accuracy and quick recognition time, the SVM model trained on the feature vectors extracted from the ResNet model is the appropriate method for the proposed system. The recognition accuracy reaches more than 98% on the well-matched testing set and more than 82% on the high-mismatched testing set, respectively. A DUT data set of employees has been built to serve the system development and face recognition evaluation on real scenarios of the university building operation. In addition, the web-based management application was developed to provide many features such as tracking people entering the building, displaying welcome messages, and training new people into the system. Our future work is to optimize the facial recognition module and add an emotional recognition module to the system.

## ACKNOWLEDGMENT

This work was supported by The University of Danang, University of Science and Technology through the research projects with code number T2019-02-40, T2019-02-41. Specially thanks to the "Smart University towards Smart City - SmU2SmC" Research Team at UD-DUT which provided insight and expertise that greatly assisted the research in this paper.

## REFERENCES

- [1] Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. AI Now 2019 Report. New York: AI Now Institute, 2019, [https://ainowinstitute.org/AI\\_Now\\_2019\\_Report.html](https://ainowinstitute.org/AI_Now_2019_Report.html).
- [2] P. Viola and M. J. Jones, "Robust real-time face detection". In: International Journal of Computer Vision, vol. 57. no. 2. pp. 137-154. 2004.
- [3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection". In: Computer Vision and Pattern Recognition, IEEE Computer Society Conference 2005. vol. 1. pp. 886-893.
- [4] H. S. Dadi, G. K. M. Pillutla, "Improved Face Recognition Rate Using HOG Features and SVM Classifier". In: IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), Volume 11, Issue 4, Ver. I, pp. 34-44, Jul-Aug. 2016.
- [5] J. E. C. Cruz, E. H. Shigueomon, L. N. F. Guimarães, "A comparison of Haar-like, LBP and HOG approaches to concrete and asphalt runway detection in high resolution imagery". In: JCIS, Dec, 20, 2015.

- [6] Burcu Kır Savaş; Sümeyya İlkin; Yaşar Becerikli, "The realization of face detection and fullness detection in medium by using Haar Cascade Classifiers". In: 24th Signal Processing and Communication Application Conference (SIU). May,16-19,2019.
- [7] Li Cuimei; Qi Zhiliang; Jia Nan; Wu Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers". In: 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI). Oct,20-22,2017
- [8] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach". In: IEEE International Conference on Computer Vision, pp. 3676–3684, 2015.
- [9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks". In: IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, Oct 2016.
- [10] Yang Wang, Guowu Yuan D.D.S., Dong Zheng, Hao Wu, Yuanyuan Pu, and Dan Xu "Research on face detection method based on improved MTCNN network", Proc. SPIE 11179. In: Eleventh International Conference on Digital Image Processing (ICDIP 2019), 111791C (14 August 2019). doi:10.1117/12.2539617
- [11] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition". In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016
- [12] Fares Jalled, "Face Recognition Machine Vision System Using Eigenfaces". In: Computer Vision and Pattern Recognition, 2017.
- [13] Zifeng Wu, Chunhua Shen, Anton van den Hengel. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. In: Pattern Recognition 2016. Nov, 30, 2016.
- [14] Karen Simonyan, Andrew Zisserman: Very Deep Convolution Networks for Large-Scale Image Recognition. In: ICLR 2015. Apr, 10, 2015.
- [15] A. Canziani, E. Culurciello, A. Paszke. An Analysis of Deep Neural Network Models for Practical Applications. In: ICLR 2017.
- [16] A. Japa and Y. Shi, "Parallelizing the Bounded K-Nearest Neighbors Algorithm for Distributed Computing Systems". In: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0038-0045, 2020. doi: 10.1109/CCWC47524.2020.9031198.
- [17] E. Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. Thirion, B. Grisel, O. Blondel, M. Prettenhofer, P. Weiss, R. Dubourg, V. Vanderplas, J. Passos, A. Cournapeau, D. Brucher, M. Perrot, M. Duchesnay: Scikit-learn: Machine Learning in Python. In: Journal of Machine Learning Research, vol 12, pp. 2825-2830, 2011.
- [18] Cambridge P. J. Phillips, H. Wechsler, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 16(5):295–306, 1998.

# Saliency Prediction for 360-degree Video

Chuong H. Vo

University of Science and Technology  
University of Danang  
Danang, Vietnam  
vhchuong1997@gmail.com

Jui-Chiu Chiang

Department of Electrical Engineering  
National Chung Cheng University  
Chia-Yi, Taiwan  
rachel@ccu.edu.tw

Duy H. Le

University of Science and Technology  
University of Danang  
Danang, Vietnam  
lhduy@dut.udn.vn

Thu T.A. Nguyen

University of Science and Technology  
University of Danang  
Danang, Vietnam  
ntathu@dut.udn.vn

Tuan V. Pham

University of Science and Technology  
University of Danang  
Danang, Vietnam  
pvtuan@dut.udn.vn

**Abstract**— Saliency detection simulates human's perception in locating crucial regions, enabling further processing for many practical applications. Even though saliency prediction for conventional 2D images and videos have been well developed, prediction on 360° contents is still challenging. For each pixel in the equirectangular frame, there will be corresponding surrounding pixels according to their spherical coordinate. Therefore, the conventional convolution method may induce certain inaccuracy in attempt to simulate humans perceive the surrounding environment. This paper proposes a novel spherical convolutional network concentrating on 360° video saliency prediction in which the kernel is defined as a spherical cap. In the process of convolution, instead of using neighboring pixels with regular relationship in the equirectangular projection coordinate, the convolutional patches will be changed to preserve the spherical perspective of the spherical signal. Our model is trained and tested on the dataset including 104 360° videos that comprise dynamic sporty content. The proposed spherical convolutional network is evaluated by Pearson correlation coefficient (CC) and Kullback-Leibler divergence (KLD). Our experiments show the efficiency of our proposed spherical convolution method's application in 360° video saliency detection utilizing spherical U-net model. Further analysis on the proposed system have been presented in this study.

**Keywords**— Video saliency detection, 360° videos, Panorama videos, Spherical convolution, Convolutional neural network

## I. INTRODUCTION

Saliency detection, also known as visual attention prediction, is the process of identifying the objects or regions that tend to draw human's attention in each scene. This task enables the machine vision system to simulate human's perception in locating crucial regions for further processing or maybe even analyzing the unimportant surrounding environment. Until now, there have been a huge number of applications utilizing saliency detection such as Automatic foveation for video compression [1], automatic image retargeting [2], information positioning scheme [3].

The most common method of researching about Saliency detection involves arranging participants to gaze at images or videos with a limited field-of-view (FoV) in a limited time interval with an eye tracker adopted for recording their eye fixations. Nevertheless, differing from this technique, human beings perceive the real world in an active manner. They usually rotate their heads in order to fully explore the

environment, therefore forming an omnidirectional understanding of the panorama scene.

Even though saliency prediction for conventional 2D images and videos have been well developed, prediction on 360° contents is still challenging. For each pixel in the equirectangular frame, there will be corresponding surrounding pixels according to their spherical coordinate. Therefore, the conventional convolution method may induce certain inaccuracy in attempt to simulate humans perceive the surrounding environment. For this reason, there needs to be a convolutional method that can mitigate the imprecision resulted by the distortion of ERP. We propose a novel spherical convolutional network concentrating on 360° video saliency prediction in which the kernel is defined as a spherical cap. In the process of convolution, instead of using neighboring pixels with regular relationship in the equirectangular projection coordinate, the convolutional patches will be changed in order to preserve the spherical perspective of the spherical signal.

## II. RELATED WORK

### A. Convolutional Neural Networks applied in Spherical Data

Through the history of machine learning, convolutional neural network (CNN) has proved its efficacy in computer vision relating applications in which the conventional data being processed are perspective images. Even though several approaches to 2D images and videos have been developed, there is still limited number of studies on 360° contents.

Up till now, the most common method in processing spherical signals is projecting them onto perspective views such as equirectangular projection (ERP), cube map projection (CMP), equi-angular cube map projection (EAC), etc... This way of approaching spherical content appears to be inefficient and feeble. Apportioning the spherical images into small partitions and projecting them with local perspective projection can result in high computational cost since saliency detection would be performed on each tile. Employing perspective-based saliency detection in processing panorama contents also raises up a certain problem that the geometric distortion caused by equirectangular projection can make numerous salient objects omitted by ordinary 2D approach. As spherical data are completely different from 2D contents, the method of analyzing them must be different or maybe translated in some respects in order to preserve their unique characteristics. In the attempt of performing CNN on spherical

data, V. Sitzmann et al. [4] projected spherical images on to a plane using equirectangular projection and performed standard CNN on them in a similar way to processing ordinary 2D images. Nonetheless, ERP would possess some certain distortion increasing as the data are located nearer to south and north poles. According to the parameter sharing property of CNN, the size and parameter of the kernel should be the same in every step of convolution. Correspondingly, in CNN employed at spherical signals, the kernels in the same size are applied throughout the sphere. The problem arises that for each specific spherical coordinate ( $\theta$ ,  $\phi$ ), a patch corresponds to a region with different shape from other coordinates. As a result, using standard CNN with the shared kernel loses its perceptual meaning in spherical signal aspect.

In order to deal with the mention issue, Yu-Chuan Su and Kristen Grauman [5] proposed to change the shape of the convolution kernel to make up for the distortion caused by the expansion of data near the poles of the sphere. However, the filters' shape in this proposal is dependent on the longitude of the patches on the spherical coordinate, so the kernels are not shared among all positions instigating the expensive computational and storage costs. In another attempt, researchers in [6] projected spherical images repeatedly onto all its tangent planes and conducted conventional convolution on the newly yielded planar images. This solution can improve accuracy in convolution, but it induces expensive computational cost. In addition, the disjoint among the tangent planes make it impossible to share the intermediate representation for higher layer convolution. Although these approaches employ tailored convolutional networks for the spherical panoramic data, the primitive intuition of mimicking human beings' way of perceiving real 360° contents was still not thoroughly looked into. In practice, when human explores the 360° contents, the human brain adopts the same mechanism to distinguish salient objects for every distinct view angle or field of view. In other terms, in the attempt of using the CNN for 360° video saliency detection, we should keep the same kernel size and parameters for all the convolution operation on different view angle or FOV.

### B. Saliency Detection implemented in videos

Thus far, though numerous efforts gone into studying conventional video saliency detection, including hand-crafted features-based methods [7][8][9] and deep learning-based methods [10][11][12], the researches in 360° videos saliency detection have been leaving several unconsidered aspects. K. Ruhland et al. [13] pioneered in 360° data saliency detection but the input data used in this research are in static image format. In the attempt to comprehend human's perception in 360° videos involving dynamic features like sports, H. Hu, Y. Lin et al. [14] proposed to locate the salient objects by managing CNN to straightforwardly process equirectangular projected images. Despite attaining acceptable outcome, this method did not take the distortion resulted by equirectangular projection into account, therefore negatively affected the accuracy. Moreover, this approach deviated from the nature of human beings' visual attention as it utilized manually labeled salient objects. Zhang et al. [15] considered about the distortion in ERP projection and attempted to compensate for it by proposing to define the shape of the kernel is a spherical cap to indicate that the kernel is shared among the convolution layer. This approach is vivid, but it was memory consuming since it created several shapes of kernel in the process of convolution.

## III. PROPOSED METHOD

There needs to be a more accurate approach in order to thoroughly mimic real human's vision mechanism. This motivates us to develop a new way of executing spherical convolutional network that could compensate for the distortions caused by ERP projection by altering the inputs' perception before performing convolution on them. In this section, we introduce our method of performing spherical convolution on panorama contents. Noting that the 2D convolution operation notation in deep learning generally implies the correlation process in mathematic.

### A. Implement Details

In the work of Zhang et al [15], the author proposed to define the shape of the kernel is a spherical cap to indicate that the kernel is shared among the convolution layer. By changing the shape of the ERP kernel according to the kernel's coordinate, the perspective meaning of the kernel is preserved. However, the complexity of changing the kernel continuously in the process of convolution is very high.

We also define the spherical convolutional kernel as a spherical cap. But instead of changing the kernel shape, we change the perspective of the input image. During applying ERP projection from spherical image to panorama image, there is distortion increasing for the pixels nearer to the north and south poles. Therefore, we need to preprocess the input image before performing convolution.

In this case, we choose the kernel used in convolutional layer is 3x3 kernel. In the convolution process, the kernel will be elementwisely multiplied with each 3x3 patch of the image. Our intention is to change the values in each 3x3 patch so that the perspective of the spherical image can be preserved during the convolution. We refer to the spherical coordinate system and the ERP projection defined from J-VET conference [16] that are shown in Fig. 2.

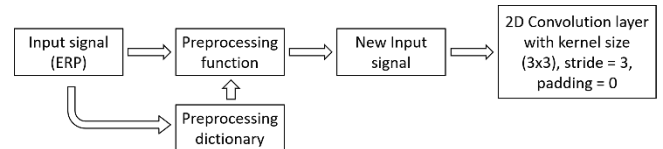


Fig. 1. The framework of processing signal in the spherical convolutional layer

The framework of processing signal in the spherical convolutional layer is illustrated in Fig. 1. After being processed by the preprocessing function, the new input signal will have the size 3 times as big as the original input signal. To preserve the perspective of the image, we perform the convolution with stride=3 and padding=0. The output of the convolution will have the same size of the input and the number of channels we decide in CNN structure.

1) *Forming spherical cap's surrounding coordinate:* The procedure consists of 6 steps:

- Step 1: Locate the surrounding points spherical cap at the north pole
- Step 2: Convert the located points' spherical coordinate into Cartesian coordinate
- Step 3: Determine the spherical coordinate of the corresponding center pixel on the ERP image

- Step 4: Rotate the original points around z axis
- Step 5: Rotate the points around x axis

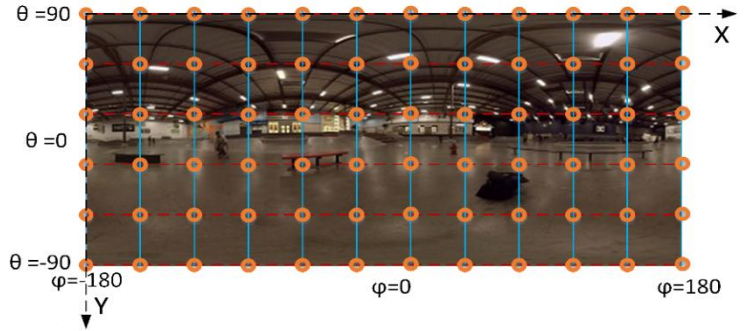
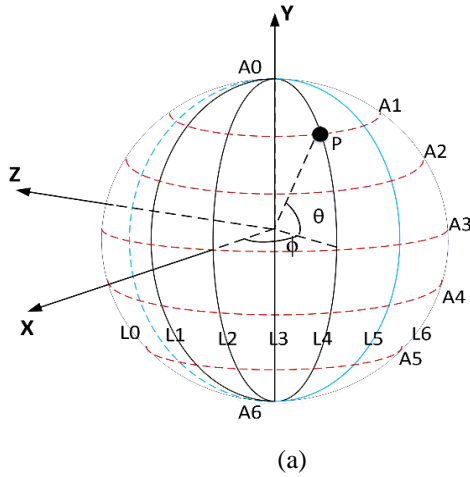


Fig. 2. 3D XYZ coordinate definition, A3 is the equator [16]; (b) Sampling coordinate definition in (X, Y) plane in our work

a) Locating the surrounding points spherical cap at the north pole.

- The original spherical cap has the center point located at the north pole  $(\frac{\pi}{2}, 0)$ .
- The angular radius of the spherical cap is chosen as  $\theta_r$ .
- The surrounding points will have the coordinates illustrated in Fig. 3.

$(\frac{\pi}{2} - \theta_r, -\frac{3\pi}{4})$	$(\frac{\pi}{2} - \theta_r, \pi)$	$(\frac{\pi}{2} - \theta_r, \frac{3\pi}{4})$
$(\frac{\pi}{2} - \theta_r, -\frac{\pi}{2})$	$(\frac{\pi}{2}, 0)$	$(\frac{\pi}{2} - \theta_r, \frac{\pi}{2})$
$(\frac{\pi}{2} - \theta_r, -\frac{\pi}{4})$	$(\frac{\pi}{2} - \theta_r, 0)$	$(\frac{\pi}{2} - \theta_r, \frac{\pi}{4})$

Fig. 3. The spherical coordinate of the spherical cap's surrounding points

- We save the spherical coordinate of the original points in a matrix as followed:

$$\begin{bmatrix} \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r & \frac{\pi}{2} - \theta_r \\ 0 & \frac{\pi}{4} & \frac{\pi}{2} & \frac{3\pi}{4} & \pi & \frac{5\pi}{4} & \frac{3\pi}{2} & \frac{7\pi}{4} \end{bmatrix}$$

b) Converting the located points' spherical coordinate into Cartesian coordinate.

- Using the formulas obtained from J-VET conference [16]:

$$X = \cos \theta \cos \phi \quad (1)$$

$$Y = \sin \theta \quad (2)$$

$$Z = -\cos \theta \sin \phi \quad (3)$$

- We obtain the Cartesian coordinate of the original points.

- Step 6: Convert the rotated points' Cartesian coordinate to ERP coordinate

- The Cartesian coordinate of the original points is stored as:

$$\begin{bmatrix} x_0 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ y_0 & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 \\ z_0 & z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 \end{bmatrix}$$

c) Determining the spherical coordinate of the corresponding center pixel on the ERP image.

- We denote x, y as the ERP coordinate
- We use the formulas:

$$\phi = x \frac{2\pi}{W-1} - \pi \quad (4)$$

$$\theta = y \frac{-\pi}{H-1} + \frac{\pi}{2} \quad (5)$$

to determine the spherical coordinate of the corresponding center pixel on the ERP image.

d) Rotating the original points around Z axis.

- We multiply each coordinate set with the rotation matrix:

$$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix}_{new} = \begin{bmatrix} \cos \Delta\theta & -\sin \Delta\theta & 0 \\ \sin \Delta\theta & \cos \Delta\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} \quad (\Delta\theta = \theta - \frac{\pi}{2}) \quad (6)$$

e) Rotating the points around X axis.

- We multiply each coordinate set with the rotation matrix:

$$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix}_{new} = \begin{bmatrix} \cos \Delta\phi & 0 & \sin \Delta\phi \\ 0 & 1 & 0 \\ -\sin \Delta\phi & 0 & \cos \Delta\phi \end{bmatrix} \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} \quad (\Delta\phi = \phi) \quad (7)$$

f) Converting the rotated points' Cartesian coordinate to ERP coordinate.

- We use the formulas obtained from J-VET conference [16]:

$$\phi = \tan^{-1}(-Z/X) \quad (8)$$



$$\theta = \sin^{-1}(Y/(X^2+Y^2+Z^2)^{1/2}) \quad (9)$$

to convert the rotated points' Cartesian coordinate to spherical coordinate.

- Then, we use the formulas

$$X = \varphi \frac{W-1}{2\pi} + \frac{W+1}{2} \quad (10)$$

$$Y = \theta \frac{1-H}{\pi} + \frac{H+1}{2} \quad (11)$$

to convert the spherical coordinate to ERP coordinate.

- Finally, we obtain a set of surrounding coordinates stored as:

$$\begin{bmatrix} X_0 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\ Y_0 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 \end{bmatrix}$$

## 2) Changing the perspective of the image.

For a conventional picture with resolution (240x480), we firstly take out 240x480 = 1,152,000 pixels and put them as the center pixel of 1,152,000 (3x3) patches. After that, we replace all the surrounding pixels of these patches with the corresponding surrounding pixels according to their coordinate in spherical domain. Next, we put these patches together in order to form a new image with resolution: 720x1440. The main framework is illustrated in Fig. 4.

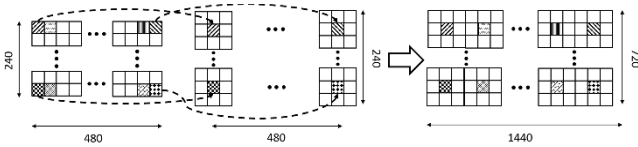


Fig. 4. Framework of the preprocessing function

For an example 3x3 patch from the image The central coordinate is: (0,0). By searching the (0,0) coordinate in the dictionary file, we acknowledge that the corresponding surrounding coordinates are:

$$\begin{bmatrix} 479 & 59 & 119 & 179 & 239 & 299 & 359 & 419 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The patch's surrounding pixels will be replaced by the pixels with the other pixels on the image having the surrounding coordinates in the matrix obtained above. This process is illustrated in Fig. 5.

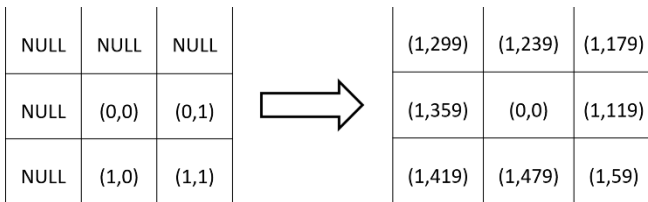


Fig. 5. The process of replacing the surrounding pixels in the 3x3 patch

After changing each the patch in the image, we can put them together to create a new image with each size is 3 times larger than the original one. The new image will become the input of the next convolutional layer.

## B. 360° Video Saliency prediction spherical U-Net

The model utilized for saliency map inference in this work is Spherical U-net model inspired by the accomplishment of Olaf Ronneberger et al [17] and Ziheng Zhang et al [15]. The model consists of a contracting path (left side) and an expansive path (right side). The contracting path bears resemblance with conventional convolutional network comprising of three convolutional layers with kernel size is 3x3, stride = 3 and padding = 0 followed by 3 Rectified linear unit (ReLU) and 3 2x2 spherical max pooling in charge of down sampling correspondingly. The expansive path consists of 3 spherical “up-convolutional” layers followed by 3 ReLU and 3 up-sampling layers correspondingly. Especially, the inputs of the 3 spherical “up-convolutional” layers are the concatenation of the convolutional results of their prior layer and the counterpart layer that possess the same metrics from the contracting path.

In our new model, we increased the size of the input from 150x300 to 240x480 in order to obtain more information from the input images. Before each convolutional layer, there is a preprocessing step that changes the perspective of the image. After the preprocessing step, the input becomes a new input with size 3 times bigger than the former one. By using the kernel size 3x3 and stride = 3 in the convolutional layer, we propose to imitate the process of spherical convolutional layer only by changing the perspective of the image instead of changing the shape of the kernel according to its position. By implementing this method, the computational cost and memory cost can be decreased. The input is the concatenation of the 3-channel image (frame) and the saliency map of its previous frame. Therefore, the input of the model has 4 channels. For each input's size of the convolution layer, there is a different radius angle of the spherical cap determined. These values are shown in table I.

TABLE I. THE RADIUS OF THE KERNEL FOR EACH IMAGE SIZE

Image size	Radius angle
240 x 480	1.1°
120 x 240	2.2°
60 x 120	4.4°
30 x 60	8.9°

Fig. 6. illustrates the new spherical U-net model in graphical manner.

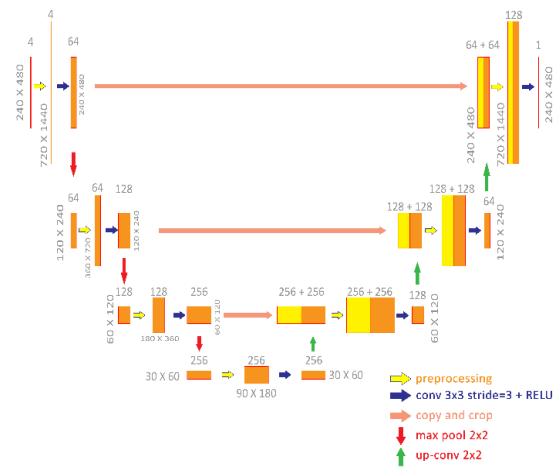


Fig. 6. Spherical U-net Convolutional network

Table II shows the architecture of the CNN in details.

TABLE II. THE ARCHITECTURE OF THE CNN

Layer	Operation	Input size	Input source	Output size
0	Input	-	-	$4 \times 240 \times 480$
1	Preprocessing	$4 \times 240 \times 480$	Layer 0	$4 \times 720 \times 1440$
2	Convolution	$4 \times 720 \times 1440$	Layer 1	$64 \times 240 \times 480$
3	Max pooling	$64 \times 240 \times 480$	Layer 2	$64 \times 120 \times 240$
4	Preprocessing	$64 \times 120 \times 240$	Layer 3	$64 \times 360 \times 720$
5	Convolution	$64 \times 360 \times 720$	Layer 4	$128 \times 120 \times 240$
6	Max pooling	$128 \times 120 \times 240$	Layer 5	$128 \times 60 \times 120$
7	Preprocessing	$128 \times 60 \times 120$	Layer 6	$128 \times 180 \times 360$
8	Convolution	$128 \times 180 \times 360$	Layer 7	$256 \times 60 \times 120$
9	Max pooling	$256 \times 60 \times 120$	Layer 8	$256 \times 30 \times 60$
10	Preprocessing	$256 \times 30 \times 60$	Layer 9	$256 \times 90 \times 180$
11	Convolution	$256 \times 90 \times 180$	Layer 10	$256 \times 30 \times 60$
12	Up-sampling	$256 \times 30 \times 60$	Layer 11	$256 \times 60 \times 120$
13	Preprocessing	$(256 + 256) \times 60 \times 120$	Layer 12 & 8	$512 \times 180 \times 360$
14	Convolution	$512 \times 180 \times 360$	Layer 13	$128 \times 60 \times 120$
15	Up-sampling	$128 \times 60 \times 120$	Layer 14	$128 \times 120 \times 240$
16	Preprocessing	$(128 + 128) \times 120 \times 240$	Layer 15 & 5	$256 \times 360 \times 720$
17	Convolution	$256 \times 360 \times 720$	Layer 16	$64 \times 120 \times 240$
18	Up-sampling	$64 \times 120 \times 240$	Layer 17	$64 \times 240 \times 480$
19	Preprocessing	$(64 + 64) \times 240 \times 480$	Layer 18 & 2	$128 \times 720 \times 1440$
20	Convolution	$128 \times 720 \times 1440$	Layer 19	$1 \times 240 \times 480$

The architecture of the CNN is shown in table II.

We utilized the loss function defined by Zhang et. al [15] for training process:

$$\mathcal{L} = \frac{1}{n} \sum_{k=1}^n \sum_{\theta=0, \phi=0}^{\theta, \phi} \omega_{\theta, \phi} (S_{\theta, \phi}^{(k)} - \hat{S}_{\theta, \phi}^{(k)})^2 \quad (12)$$

#### IV. DATASET AND TRAINING

##### A. Dataset

We utilized the video saliency dataset created in Ziheng Zhang's work [15] that entails 104 360° videos perceived by 27 viewers. The videos from this dataset were collected from Sports-360 dataset [14]. After eliminating the videos possessing the length less than 20 seconds, the remained videos are chosen for this dataset. The content of the dataset consists of 5 dynamic sports including basketball, BMX, dancing, skateboarding and parkour). The creator of this dataset used an HTC VIVE HMD and a '7invensun a-Glass' eye tracker for extracting the eye gazing data of the videos' viewers. There were 27 volunteers aged between 20 and 24 years old recruited in the dataset creating experiments. All 104 videos are divided into 3 sessions and each session contains 35 360° videos. Volunteers are requested to gaze on 360° videos starting at a fixed location ( $\theta = 90$ ,  $\phi = 180$ ) in random orders.

In order to perform experiments in our work, 80 videos are chosen randomly to be the training data while the other 24 videos are assigned for testing the model.

##### B. Training the spherical CNN.

We implemented the spherical U-Net using the PyTorch framework. We trained our network with the following hyperparameters setting mini-batch size (2), learning rate ( $1e-4$ ), momentum (0.9), weight decay (0.00005), number of epochs (20). During the process of training, we evaluated the outputs of the model after every epoch trained and came into conclusion that the epoch 20 had the best performance among all other cases.

Our proposed 360° video saliency prediction model is trained on 1 NVIDIA 1080Ti GPU. We measure the typical training time for each image batch. The average running time of our model is 1.78 s/iteration. The spherical U-Net illustrated in Fig. 6, possesses about 1.7 M parameters. It takes approximately 7 hours in order to train the model with the total number of iterations amount to 14600.

#### V. RESULTS AND EVALUATION

##### A. Result of the saliency prediction model

The Qualitative results of the saliency prediction model are shown in Fig.7.

##### B. Performance Evaluation

*Metrics.* Since our model was applied in 360° videos case, we employed the metrics including CC and KLD established in [18] to calculate the errors between the predicted output saliency maps from and the ground truth.

After training the spherical convolutional network applied in 360° video saliency prediction, we performed the evaluation on the model. We compare our method with the baselines in table III.

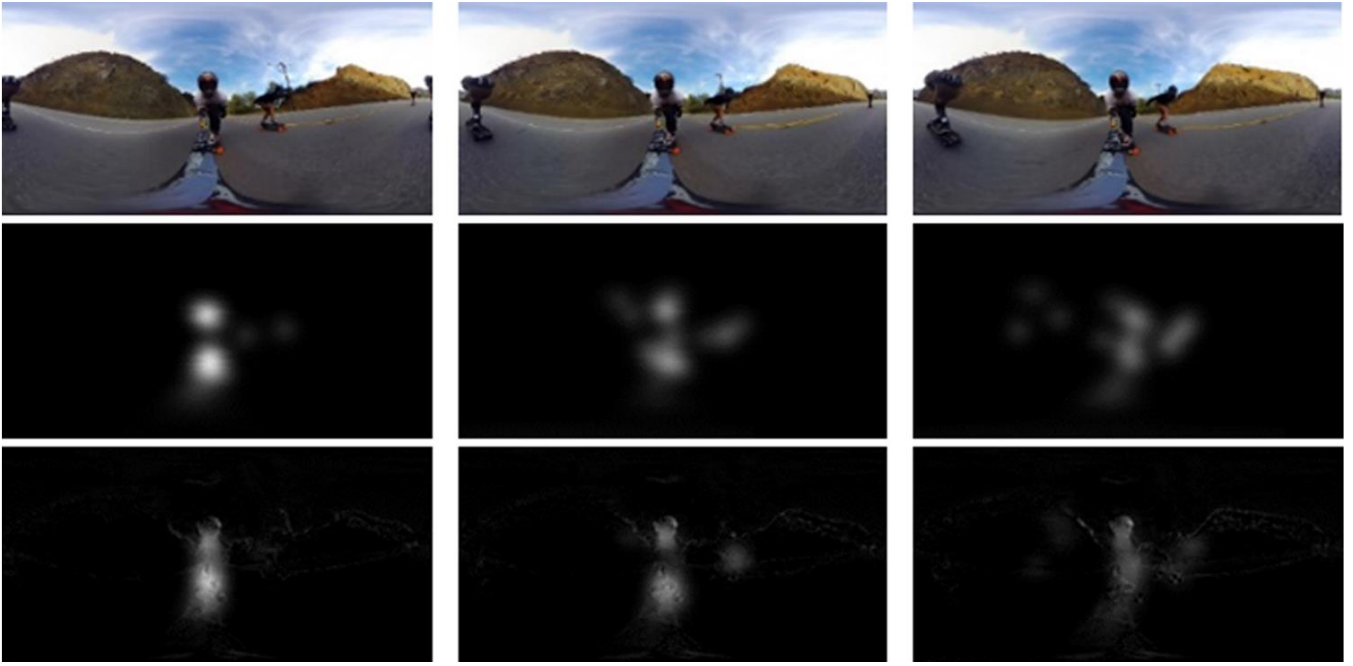


Fig. 7. Qualitative results of the saliency prediction model. The three rows show the consecutive RGB frames, their ground truth saliency maps, and the predicted saliency maps

**Baselines.** We compare our proposed spherical U-Net with the following methods including Ziheng Zhang et. al [15] and U-Net without spherical convolution which has similar structure as our model except that the convolutional layers are conventional convolution. We trained and tested the baselines with the same configuration of our proposed model on the same dataset.

As we can see, our proposed model outperformed the conventional U-Net model while it had the approximate performance compared to the other baseline. Acknowledging that in [15], the author resampled the kernel based on each coordinate of it when it is used for convolution. Therefore, this method consumed up to 84 GB of GPU memory by storing several kernels while our model only occupies 7.5 GB of GPU memory by storing the dictionary of the surrounding coordinates in the outer folder, it can be inferred that our model could have approximate performance compared to [15] but with lower memory cost of GPU's memory consumption.

TABLE III. THE PERFORMANCE COMPARISON OF OTHER METHODS WITH OUR SPHERICAL U-NET ON THE SAME VIDEO SALIENCY DATASET.

	Ziheng Zhang et al. [15]	Our proposed model	U-Net w.o. spherical convolution
CC	0.8432	0.8233	0.6971
KLD	2.2969	4.3199	7.7745

## VI. CONCLUSION

In this work, the saliency prediction in dynamic 360° video is developed. In order to achieve the project's goal, we proposed a new type of spherical CNN where the perception of the spherical input is preserved during the convolution process. Since the 360° videos are mainly stored with ERP format, we apply spherical CNN to the panorama case where the perception of each frame is altered in order to preserve the spherical convolution meaning on panorama. Thereafter, we apply the spherical convolution method into the spherical U-Net model aiming to predict 360° video saliency. Our experiments showed the efficiency of our proposed spherical

convolution method's application in 360° video saliency detection utilizing spherical U-Net model with the value of CC reaching 0.8233 and the KLD achieving 4.3199.

In order to improve the accuracy of the saliency prediction model for 360° videos, we would like to inspect in how to perform up sampling considering 360° content in the future.

## ACKNOWLEDGMENT

This work was carried out under capstone project collaboration between by Faculty of Advanced Science and Technology, The University of Danang - University of Science (UD-DUT) and Technology and the College of Engineering, National Chung Cheng University, Taiwan, Republic of China. This work is technically supported by the College of Engineering, National Chung Cheng University, Taiwan, Republic of China. Specially thanks to the SmU2SmC Program at UD-DUT which provided insight and expertise that greatly assisted the research in this paper.

## REFERENCES

- [1] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," in *IEEE Transactions on Image Processing*, vol. 13, no. 10, Oct. 2004, pp. 1304-1318.
- [2] Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B.: "Automatic image retargeting". In: Proceedings of the 4th international conference on Mobile and ubiquitous multimedia (MUM '05), ACM, New York, NY, USA, 2005, pp. 59-68
- [3] Chang, M.M.L., Ong, S.K., Nee, A.Y.C.: "Automatic information positioning scheme in AR-assisted maintenance based on visual saliency". In: International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Springer 2016, pp. 453-462
- [4] V. Sitzmann et al., "Saliency in VR: How Do People Explore Virtual Environments?" in *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633-1642, April 2018.
- [5] Su, Y.C., Grauman, K.: "Learning spherical convolution for fast features from 360 imagery". In: Advances in Neural Information Processing Systems, pp. 529-539, 2017
- [6] Y. Su and K. Grauman, "Making 360° Video Watchable in 2D: Learning Videography for Click Free Viewing," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 1368-1376.

- [7] Zhong, S.h., Liu, Y., Ren, F., Zhang, J., Ren, T., "Video saliency detection via dynamic consistent spatio-temporal attention modelling". In: Twenty-Seventh AAAI Conference on Artificial Intelligence. (2013) 1063–1069
- [8] F. Zhou, S. B. Kang and M. F. Cohen, "Time-Mapping Using Space-Time Saliency," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 3358-3365.
- [9] Itti, L., Dhavale, N., Pighin, F.: Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI. Volume 5200., International Society for Optics and Photonics (2003), pp. 64–79
- [10] Bak, C., Erdem, A., Erdem, E.: Two-stream convolutional networks for dynamic saliency prediction. arXiv preprint arXiv:1607.04730 (2016)
- [11] W. Wang, J. Shen and L. Shao, "Video Salient Object Detection via Fully Convolutional Networks," in *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38-49, Jan. 2018.
- [12] S. Chaabouni, J. Benois-Pineau and C. Ben Amar, "Transfer learning with deep networks for saliency prediction in natural video," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 1604-1608.
- [13] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. 2015. "A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception". *Journal Computer Graphics Forum*. Volume 34 Issue 6, September 2015, pp. 299-326.
- [14] H. Hu, Y. Lin, M. Liu, H. Cheng, Y. Chang and M. Sun, "Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Videos," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 1396-1405.
- [15] Zhang, Z., Xu, Y., Yu, J., & Gao, S. (2018). "Saliency Detection in 360° Videos". *The European Conference on Computer Vision (ECCV)* 2018, 2018, pp. 488-503
- [16] Yan Ye, Jill Boyce: "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib Version 8". In: Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Macau, CN, 3–12 Oct. 2018.
- [17] Ronneberger, O., Fischer, P., Brox, T. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241
- [18] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba and F. Durand, "What Do Different Evaluation Metrics Tell Us About Saliency Models?" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740-757, 1 March 2019.

# Heart Rate Estimation Based on Facial Image Sequence

Dao Q. Le

Faculty of Advanced  
Science and Technology  
(FAST)

The University of Danang -  
University of Science and  
Technology  
Danang, Vietnam  
quangdao215@gmail.com

Wen-Nung Lie

Department of Electrical  
Engineering, Center for  
Innovative Research on  
Aging Society (CIRAS),  
National Chung Cheng  
University

Chia-Yi, Taiwan  
ieewnl@ccu.edu.tw

Quynh Nguyen Quang Nhu

Faculty of Advanced  
Science and Technology  
(FAST)

The University of Danang -  
University of Science and  
Technology  
Danang, Vietnam  
nqnquynh@dut.udn.vn

Thu T.A. Nguyen

Faculty of Advanced  
Science and Technology  
(FAST)

The University of Danang -  
University of Science and  
Technology  
Danang, Vietnam  
ntathu@dut.udn.vn

**Abstract**— Recently, remotely obtaining the photoplethysmogram (PPG) signal to estimate the blood volume pulse or the heart rate of a human has become a research topic which gains increasing attention. In contrast to the longstanding contact methods (e.g., electrocardiogram (ECG)), the remote PPG methods can tackle the same task with superior convenience and fewer physical constraints. It has been proved in studies that PPG signal affects the change of color intensity on some body parts such as the face and wrist. Leveraging this, we propose a new method that uses Intel RealSense camera to capture RGB facial videos of human subjects to estimate the heart rate in a short segment of time. By combining a series of image and signal processing techniques, e.g., face detection, facial segmentation, Independent Component Analysis (ICA), filtering, Fast Fourier Transform (FFT), and a new proposed Automatic Component Selection (ACS) algorithm, we are able to accurately estimate the heart rate from the human facial video. Our method works well with slight head motion. The time length of the required facial video is also greatly reduced to about 10 seconds (traditionally, 30~60 seconds). By experiments, we achieved a root mean square error of 3.41 beats per minute (bpm) for 10-seconds RGB videos. This proved the robustness of our new defined region of interest (ROI) for the inputs and proposed ACS can provide. In our future work, shorter video clips (e.g., less than 5 seconds) and tolerance of larger head movements would be achieved so that our system can be well-applied in realistic life.

**Keywords**—Independent Component Analysis (ICA), remote PPG, heart rate estimation, facial video.

## I. INTRODUCTION

Heart rate (HR) reveals a process of rhythmic oscillation of the vessels, corresponding to the contractions of the heart. It is one of the most vital indicators that help monitor a person's health condition. It represents the heart's frequency of bumping the blood over the circulatory system, so one can understand and analyze some syndromes of the heart based on the tendency of the time series data. To accurately measure the HR, one can use special devices like wrist pulsometer, oximeter or electrocardiograph (ECG) monitor. Those mentioned methods imply contact-based measurements, but people having burnt skin or lethally contagious disease find it difficult or even impossible to have their HRs monitored. Therefore, a method of remotely extracting the HR signal from a region on human body is demanded.

Contactless estimation of HR has been investigated using different techniques and devices under various conditions. Related techniques can be divided into 3 main categories which are Doppler radar, thermal camera, and traditional camera. The studies on using Doppler radar [1], [2], [3], to track the HR efficiently via the subtle motions of the chest have achieved good accuracies. However, the equipment's cost and substantial constraints, e.g., fixed radar position and static chest make it difficult to apply to real life situations. Thermal-based methods [4], [5], infer the HR from the variation of temperatures in facial region. Similar to radar-based methods, thermal camera suffers from its expensive price. Methods based on traditional RGB cameras [6], [7], [8], to capture human body's intensity variations are however much more affordable. Though the variations are exceedingly subtle and practically mixed with other fluctuations caused by artifacts, they can be amplified/processed and observed by signal processing techniques to extract photoplethysmogram (PPG) information, a noninvasive means of sensing the cardiovascular blood volume pulse (BVP). During the cardiac cycle, volumetric change in the facial blood vessels modifies the path of the incident ambient light such that subsequent changes in the amount of reflected light will indicate the timing of cardiovascular events.

Particularly, Poh *et al.* [6] proposed an approach using Independent Component Analysis (ICA) algorithm to recover the PPG signal among the underlying RGB signals of the facial region. The largest hindrance of the ICA is the random (or, incorrect) ordering of the returned source signals. To implement an automatic system, Poh *et al.* chose to always pick the ICA component of largest peak in the power spectrum. However, this solution is susceptible in cases where noise energy dominates the PPG energy. Thus, it is difficult to obtain high accuracy in realistic scenarios.

Aiming to improve Poh's work, we propose an automatic algorithm to select the best component representing the PPG signal based on a model-matching criterion. Also in this paper, we examine three factors for remote HR estimation, they are: 1) video input length, 2) color space, and 3) ROI for signal analysis. Considering scenarios where the testees are improbable/unable to stay static in front of the camera for a long period, we explored a much shorter video input of 5 ~ 30 seconds (larger than 60 seconds in [6], [9]) to increase the diversity of applications. In this paper, we define a new ROI



which holds useful information as much as possible and keep noises far away. This new ROI is then compared with various practical ROIs, including the one used in Poh's work [6].

Section II of this paper details Poh's method [6], Section III describes our proposed algorithm, the experimental protocol and results are presented in Section IV, while Section V covers the conclusion of our work.

## II. RELATED WORK

The framework of Poh's method includes steps of: 1) video input, 2) face detection and ROI localization, 3) pre-processing, 4) ICA, 5) filtering and interpolation, and 6) HR calculation.

First, a regular RGB webcam is used to capture the video input. To locate the specific region containing necessary PPG data, a Haar Cascade based face detection [10] algorithm is performed to identify a box containing the face. After that, an ROI is then defined with a full height and a 60% width of the box. Their resulting ROI actually holds the background and some noisy parts (e.g., eyes and mouth which will contribute a great deal of artifacts when a person is talking or blinking). Additionally, it is probable that the forehead is covered by hair, which certainly prevents an accurate HR estimation.

Within the ROI, RGB signals are spatially averaged over all pixels to form raw time series (or, traces), which are then normalized, and detrended [11] (smoothed) as pre-processing. ICA is then applied to each trace to uncover the independent PPG signal by picking up the component that has the largest energy peak in the frequency domain. The assumption of dominance of PPG signal is not always true due to diversities of artifacts in realistic cases, which might overwhelm the subtle variations due to blood volume changes.

The separated PPG signal is then smoothed using a 5-point moving average filter and a bandpass filter with a passband of 0.7-4 Hz. This frequency range basically covers the possible heart rate range of a person, equal to 42-240 beats per minute (bpm). The filtered signal is also further up-sampled to 256 Hz by cubic spline interpolation from the original 15 Hz to refine the primary peaks. With the enhanced peaks detected, they measure the mean distance between peaks, called Inter-beat Interval (IBI), and use it to achieve the final HR estimation by  $\hat{HR} = 60/IBI$ .

## III. PROPOSED METHODOLOGY

Fig. 1 illustrates our algorithm for HR estimation. It improves Poh's work by refining ROI to reject noise and proposing an automatic component selection for HR estimation.

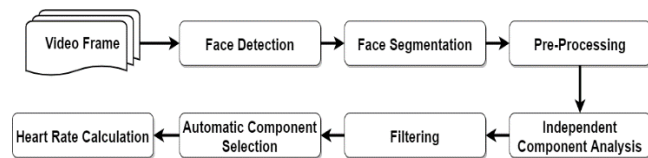


Fig. 1. Our proposed algorithm

### A. Face detection

To detect the human face from each video frame, a face detector capable of detecting Dlib's 68 facial landmarks<sup>[1]</sup> (Fig. 2(a)), in addition to a bounding box, is used. It is a method based on the feature of Histogram of Oriented

Gradients (HOG) and was trained beforehand with dataset of human faces.

The position of the face bounding box is used to align the faces in vertical direction temporally. On the other hand, coordinates of the facial landmarks at eyes and mouth are used to define excluded areas that prevent the artifacts when a person is talking or blinking.

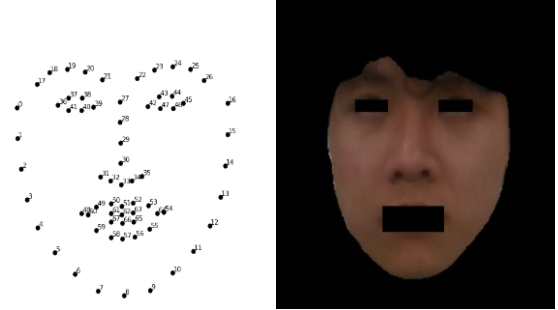


Fig. 2. The proposed ROI for HR estimation, (a) definition of 68 facial landmarks, (b) ROI mask (with skin colors) after face segmentation and eyes/mouth removal.

### B. Face segmentation

Here, provided with the face bounding box from the previous step, a pre-trained face segmentation<sup>[2]</sup> is performed to crop out non-skin regions (e.g., background and hair) and return the final ROI mask (see an example in Fig. 2(b)). This step is expected to further enhance the signal to noise ratio (SNR) for signals calculated from ROI.

The trace signal  $I[t]$  in RGB components calculated for ROI is formed via:

$$I[t] = \frac{1}{|ROI|} \sum_{(i,j) \in ROI} f_{t(i,j)}, t=1, \dots, N \quad (1)$$

where  $f_{t(i,j)}$  represents the RGB signal at time  $t$  and pixel  $(i, j)$  defined by ROI mask,  $|ROI|$  means the size (number of pixels) of ROI, and  $N$  is the number of frames in the sequence. Hence,  $I[t], t = 1, \dots, N$ , reveals the mean intensity variation over time.

### C. Pre-processing

In this step, the pre-processing includes normalization and detrending, as in [6], to help smoothen the signal and minimize the impact of large eruptions and sudden noise in general. The raw trace  $I[t]$  is normalized as follows:

$$I'[t] = \frac{I[t] - \mu}{\sigma} \quad (2)$$

where  $\mu, \sigma$  are the mean, standard deviation of  $I[t]$  and  $I'[t]$  is the normalized raw trace. From this point forward, we re-denoted the normalized raw trace as  $I[t]$  for convenience and coherence.

The normalized raw trace is then detrended by using a procedure based on a smoothness priors approach [11] with the smoothing parameter  $\lambda = 1$ .

### D. Independent Component Analysis (ICA)

With unknown sources of artifacts in  $I[t]$ , we need a processing to separate PPG signal from others. Such illuminative variation could be caused by respiration, facial

[1]: Available online on [www.github.com/davisking/dlib](http://www.github.com/davisking/dlib)

[2]: Available online on [www.github.com/nasir6/face-segmentation](http://www.github.com/nasir6/face-segmentation)

motions, facial expression changes or vestibular activities [12]. The decomposition into sub-component signals can be achieved by ICA, presuming that the PPG signal is independent of other sources. The outputs of ICA are the additive  $K$  ( $K=3$ ) subcomponents from the observation signal  $I[t]$ , denoted by  $I_{ICA,k}$ .

Grounded by [13], via a comparison between different algorithms of ICA, Joint Approximation Diagonalization of Eigen-matrices (JADE) is the most suitable algorithm for this application. The number of signals is set equal to the number of observations which is  $K=3$  (i.e., RGB), as described in [14].

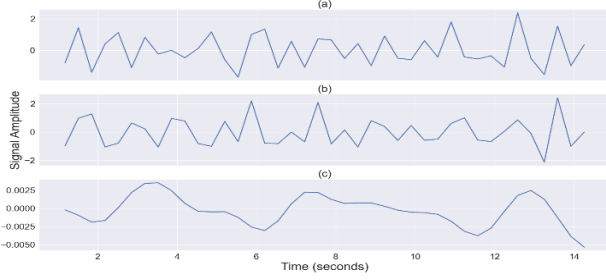


Fig. 3. One ICA component signal, (a) the signal before filtering, (b) the signal after moving-average smoothing filter, (c) the signal after moving-average smoothing and band-pass filters.

#### E. Filtering

To narrow the frequency range of our interest, a 32-point moving average filter and a Butterworth band-pass filter from 0.7 to 4 Hz are implemented (see Fig. 3). This filtering removes the secondary peaks in the ICA components  $I_{ICA,k}$  as well as avoids the irrelevant frequency ranges. Additionally, we do not interpolate the ICA components since IBI is not calculated for HR estimation.

As Poh claimed in their work [15], the maximum change of HR in 1 second apart is 12 bpm. Utilizing this, we apply another band-pass filter to focus more on the possible HR range based on previous estimation, denoted as  $i\widehat{HR}$ , with the following cutoff frequencies:

$$High\ cutoff = 1.2 \times i\widehat{HR} \quad (3)$$

$$Low\ cutoff = 0.8 \times i\widehat{HR} \quad (4)$$

Here, we denote the resulting ICA components to be  $I'_{ICA,k}$ ,  $k=1, \dots, K$ .

#### F. Automatic Component Selection

In order to overcome the randomly returned ordering of ICA components, we propose an Automatic Component Selection (ACS) procedure to automatically pick up the component presenting the wanted PPG signal. The ACS is described step by step as follows:

Step 1: Create a PPG model  $I_{model}[t]$ ,  $t=1, \dots, N$ , which is a sine wave with a frequency of 1 Hz (corresponding to a normal heart rate of 60 bpm), magnitudes running from -1 to 1 (similar to the range returned by ICA), and the length equal to the time axis of the ICA components.

Step 2: Calculate the difference signal between the PPG model signal  $I_{model}[t]$  generated in step 1 and each of the ICA component signal (here,  $I'_{ICA,k}[t]$ ,  $k=1 \sim K$ ) extracted and

processed. Denote them as  $diff_k[t] = |I_{model}[t] - I'_{ICA,k}[t]|$ ,  $k=1, \dots, K$ . An example is shown in Fig. 4.

Step 3: Detect peaks in  $diff_k[t]$  and evaluate horizontal distances between them, denoted as  $\{dist\_peaks_{k,i}\}$ , where  $i$  is the sample index.

Step 4: Calculate the skewness of samples in  $k$ -th ICA component as below:

$$\frac{1}{(N-1)} \sum_{i=1}^{N-1} (dist\_peaks_{k,i} - \overline{dist\_peaks_k})^3, \quad (5)$$

where  $\overline{dist\_peaks_k}$  is the mean of peak distances and  $N$  is the length of ICA component signal.

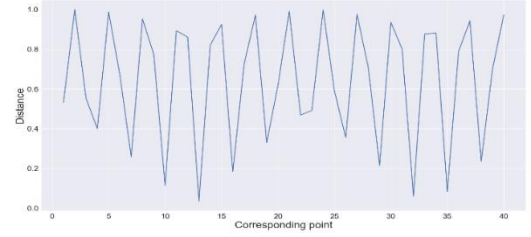


Fig. 4. An example of difference signal.

Step 5: Choose the ICA component which has the smallest absolute skewness as the PPG signal. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The smaller the absolute skewness is, the more symmetric the distribution is. Therefore, skewness is used in this paper to point out how symmetric a distribution of  $dist\_peaks_{k,i}$  is, i.e., how much the harmonic property is for  $k$ -th ICA component. An example can be seen in Fig 5 with the third ICA's component being automatically chosen as the PPG signal.

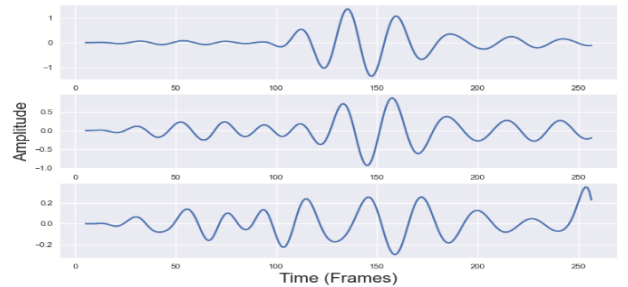


Fig. 5. Example of three inputs of the ACS, their absolute skewness values are 1.167, 0.816, and 0.378, respectively.

#### G. Heart Rate Calculation:

Instead of using IBI to calculate the HR, we utilize a more popular method, Fast Fourier Transform (FFT) to convert the PPG signal into Power Spectral Density (PSD) in its frequency domain. The frequency  $Freq_{peak}$  corresponding to the highest energy peak is then located. The estimated HR value is evaluated as:

$$\widehat{HR} = 60 \times Freq_{peak} \quad (6)$$

#### IV. EXPERIMENTAL RESULTS

Testing dataset videos were shot with a RealSense D435 camera. All of them have resolution of 640x480 pixels, frame rate of 30 per second and an average duration of 30 seconds. Videos were shot indoors, with evenly illuminative condition

as a normal working space, so that the persons' faces on videos have no glares or shades. The subjects, remained in sitting posture during the footage, are allowed to slightly move their heads.

The BIOPAC PPG100C [16] was used to measure the heart rate simultaneously as the ground truth for calculating the errors of our estimations. Having one SpO<sub>2</sub> sensor, it can be warped around the fingertip for continuously monitoring the PPG signal. The device also has an adjustable sampling rate, which was set to 500 Hz in our experiments. We also programmed an application called PPGtest on Lab Viewer. The PPGtest is connected to the device via an USB port. It is capable of monitoring the PPG in real-time and saving PPG records to a text file. For ground truth data acquisition, subjects were measured with the PPG device and text files were simultaneously recorded in accompanying with videos. In total, 9 persons with various HR range (e.g., static, after walking, or after slight exercise) participated in the recording. Similar to [6], we use Root Mean Square Error (RMSE) to evaluate our experiments.

#### A. Comparison on video length

The experiments were conducted to find out the shortest length required to reach the standard. According to ANSI/AAMI EC13:2002, an allowable readout error is no greater than  $\pm 10$  percent of the input rate, or  $\pm 5$  bpm. In our evaluation, the latter is considered as the standard. The resultant impact of the video length is illustrated in Fig. 6. It was found the least video length to meet the standard is 10 seconds.

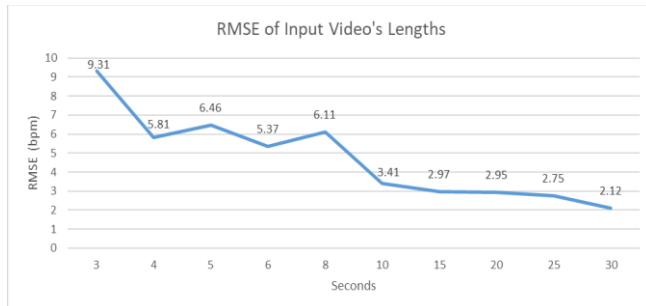


Fig. 6. RMSE of our algorithm for varying video lengths.

#### B. Comparison with prior work [6]

Another evaluation was carried out to compare the performance of our method and Poh's method [6] with 10-second video inputs. It reveals that our method has better accuracy for videos of same length at 10 seconds.

TABLE 1. RMSE OF THE PROPOSED AND POH'S METHOD

	Proposed Method	Poh's Method
Person 1	8	33.9
Person 2	3.56	0.68
Person 3	0.46	1.17
Person 4	0.99	1.87
Person 5	0.13	3.84
Person 6	1.47	3.06
Person 7	1.48	4.75
Person 8	7.53	18.65
Person 9	7.1	3.24
Average RMSE	3.41	7.91

#### C. Test with different ROIs

Besides the impact of video length, we also experimented on different ROIs and color spaces. Table 2 compares the accuracies of "proposed ROI", "Full Face", "Double Cheeks" [17] and "Poh's ROI" (see examples in Fig. 7). The videos length in this test is also selected to be 10 seconds. It was found that our ROI and RGB color space will lead to the best accuracy.



Fig. 7. Example of tested ROIs.

TABLE 2. RMSE OF DEFINED ROIS WITH SEVERAL COLOR SPACES

	RGB	HSV	HLS	CIELab	YUV	YCbCr
Our ROI	<b>3.41</b>	3.61	4.62	5.05	5.44	5.18
Full Face	5.07	7.06	8.31	5.12	6.78	5.36
Double Cheeks	5.83	6.14	5.47	5.33	6.1	6.24
Poh's ROI	6.57	8.24	5.98	5.47	6.29	7.31

## V. CONCLUSION

On the basis of the results from present study, we have demonstrated the feasibility of the proposed method to improve the estimation accuracy from Poh's backbone framework. Table 1 illustrates both individual and general accuracy where the proposed method performed better in realistic scenario. From the results in Fig. 6, the accuracy of the method is better for longer length of the video. With RMSE of 3.41 bpm, the length of 10-second video has been proved to be usable. With reduced time to sit in front of the camera, testees possibly welcome and feel comfortable with our method. This result also suggests the method's feasibility in term of performing quick vital check.

The removal of two eyes and mouth is an essential part to enhance the SNR of the input signal, as shown in Table 2. There is no certain trend among the testees when changing the color space and ROI. However, our proposed ROI generally achieved the best accuracies with all color spaces.

The conducted experiments lack of more detailed comparison such as facial expressions, activity state, pose and distance to the camera. During our experiments, we observed some false cases in face detection and face segmentation which negatively affect the results. Our finding should motivate extensive validation and continued systematic improvement of these aspects.

## REFERENCES

- [1] M. Nosrati and N. Tavassolian, "High-accuracy heart rate variability monitoring using Doppler radar based on Gaussian pulse train modeling and FTFR algorithm," *IEEE Trans. on Microwave Theory and Techniques*, vol. 66, no. 1, pp. 556-567, 2017.
- [2] Y. Nakayama, G. Sun, S. Abe and T. Matsui, "Non-contact measurement of respiratory and heart rates using a CMOS camera-equipped infrared camera for prompt infection screening at airport

- quarantine stations," in *2015 IEEE Int'l Conf. on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Shenzhen, 2015.
- [3] M. Baboli, A. Singh, N. Hafner and V. Lubecke, "Parametric study of antennas for long range Doppler radar heart rate detection," in *2012 Annual Int'l Conf. of the IEEE Engineering in Medicine and Biology Society*, San Diego, 2012.
- [4] T. R. Gault and A. A. Farag, "A fully automatic method to extract the heart rate from thermal video," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Portland, 2013.
- [5] S. L. Bennett, R. Goubran and F. Knoefel, "Adaptive eulerian video magnification methods to extract heart rate from thermal video," in *2016 IEEE Int'l Symp. on Medical Measurements and Applications (MeMeA)*, Benevento, 2016.
- [6] M.-Z. Poh, D. J. McDuff and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. on Biomedical Engineering*, vol. 58, no. 1, pp. 7-11, 2010.
- [7] K.-Z. Lee, P.-C. Hung and L.-W. Tsai, "Contact-free heart rate measurement using a camera," in *2012 Ninth Conf. on Computer and Robot Vision*, Toronto, 2012.
- [8] X. Li, J. Chen, G. Zhao and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," in *2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Columbus, 2014.
- [9] T. Kitajima, S. Choi and E. A. Y. Murakami, "Heart rate estimation based on camera image," in *2014 14th Int'l Conf. on Intelligent Systems Design and Applications*, Onikawa, 2014.
- [10] M. J. Paul Viola, "Rapid object detection using a boosted cascade of simple features," in *2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, 2001.
- [11] M. Tarvainen, P. Ranta-aho and P. Karjalainen, "An advanced detrending method with application to HRV analysis," *IEEE Trans. on Biomedical Engineering*, vol. 49, no. 2, pp. 172-175, August 2002.
- [12] G. Balakrishnan, F. Durand and J. Guttag, "Detecting pulse from head motions in video," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Portland, 2013.
- [13] V. Ostankovich, G. Prathap and I. Afanasyev, "Towards human pulse rate estimation from face video: automatic component selection and comparison of blind source separation methods," in *2018 Int'l Conf. on Intelligent Systems (IS)*, Funchal - Madeira, Sept 2018.
- [14] J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157-192, 1991.
- [15] M.-Z. Poh, D. J. McDuff and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Exp.*, vol. 18, no. 10, pp. 10762-10774, May 2010.
- [16] "BIOPAC's Pulse Plethysmogram Amplifier," BIOPAC System, Inc., [Online]. Available: <https://www.biopac.com/product/pulse-plethysmogram-amplifier/>.
- [17] R. Rahman, K. Ukai and S. Kobashi, "A filter-based method to calculate heart rate from near infrared video," in *2018 Joint 10th Int'l Conf. on Soft Computing and Intelligent Systems (SCIS) and 19th Int'l Symp. on Advanced Intelligent Systems (ISIS)*, Toyama, Japan, 2018.
- [18] H. Tan, D. Qiao and Y. Li, "Non-contact heart rate tracking using Doppler radar," in *2012 Int'l Conf. on Systems and Informatics (ICSAI2012)*, Yantai, 2012.

# Stability Analysis of an Islanded Microgrid Using Supercapacitor-based Virtual Synchronous Generator

Hong Viet Phuong Nguyen  
University of Science and  
Technology  
The University of Danang  
Danang City, Vietnam  
nhvphuong@dut.udn.vn

Van Tan Nguyen  
University of Science and  
Technology  
The University of Danang  
Danang City, Vietnam  
tan78dmbk@dut.udn.vn

Binh Nam Nguyen  
University of Science and  
Technology  
The University of Danang  
Danang City, Vietnam  
nbnam@dut.udn.vn

Thi Bich Thanh Truong  
University of Science and  
Technology  
The University of Danang  
Danang City, Vietnam  
ttbthanh@dut.udn.vn

Huu Dan Dao  
University of Science and  
Technology  
The University of Danang  
Danang City, Vietnam  
daohuudan2310@gmail.com

Quoc Cuong Le  
University of Science and  
Technology  
The University of Danang  
Danang City, Vietnam  
le.cuong.4298@gmail.com

**Abstract**— The gradually increasing number of distributed generators in microgrids cause negative affections to their dynamic stability. Moreover, dynamic performance of small synchronous generators such as diesel is usually poor due to slow response and low inertia. To overcome these drawbacks, virtual synchronous generator is considered as a promising suggestion to improve dynamic characteristics of synchronous generators in such a way of increasing microgrid inertia, especially in islanded operation. This paper analyzes the dynamic stability of microgrid including diesel generator, photovoltaic, loads and an energy storage system based on a voltage-source inverter equipped with virtual synchronous generator control strategy. Supercapacitor with fast response capability is utilized to inject demand's power from voltage-source inverter. Nonlinear mathematical models of microgrid used for this study are built in Simulink/Matlab environment. Finally, simulation results in time-domain under various operating conditions are presented to verify the effectiveness of virtual synchronous generator suggestion.

**Keywords**—distributed generator, dynamic stability, inertia, virtual synchronous generator, supercapacitor

## I. INTRODUCTION

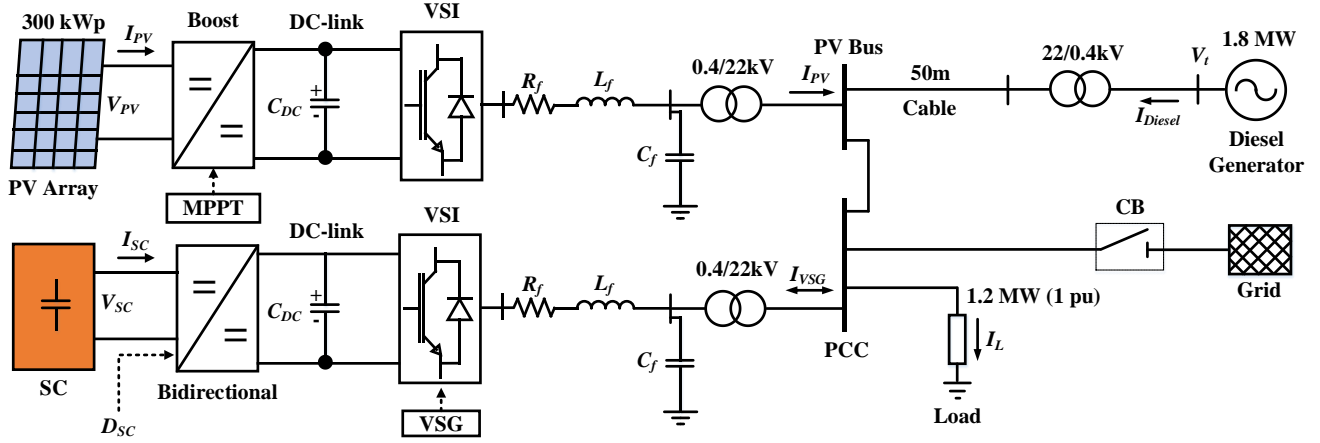
Recently, the large-scale integration of Renewable Energy Sources (RESs) into the power grid has been considered as a future solution for reducing the greenhouse effect due to the emission from the burning of fuels of the traditional power plants. Therefore, multiple traditional power plants have been gradually replaced by clean energy sources. Different from centralized conventional generators based on Synchronous Generators (SGs), most of RESs are usually made of small power ratings and directly connected to distribution grid and local loads. These are called Distributed Generators (DGs) [1], such as wind, solar generations and microturbines. To facilitate the integration of DGs in distribution systems, the concept of Microgrid (MG) is proposed [2]. Although the large penetration ratio of DGs has many benefits such as reducing voltage losses, power losses on the distribution transmission lines and improving system reliability, the DGs have a number of effects on power quality such as lack of inertia, which reduces system-wide inertia and causes frequency/voltage instability. Moreover, main generations in

MGs such as microturbines have the engine's delay responses to the rapid transient occurred in MGs. If the speed variation exceeds the permitted limit range, the operation of MGs will be shut down. A promising suggestion for MG stability is to emulate the behavior of conventional SGs into the DGs for improving inertia, stability, and reliability of MGs. The concept of Virtual Synchronous Generator (VSG) is introduced to perform the behavior of the generator primary engine [3].

Over the past few years, many virtual inertia control methods have been implemented for improving the frequency stability of the MG. Virtual inertia control is based on the proposed differential control strategies to improve the frequency stability of the system [4], [5]. [6] shows the comparison between the VSG control method and the droop method. Several studies on inertia and damping adjustment technologies have been conducted in [7]-[11]. Although the methods of virtual inertia control in [7]-[11] have shown effectiveness relied on the frequency derivative to make the change virtual inertia, such a derivative is very sensitive to noise caused by measurement. Several studies have introduced advanced control methods to determine the parameters of VSG [12], [13]. Virtual inertia constants and virtual damping coefficients are obtained by giving formulas, suitable functions of frequency deviation and voltage deviation based on Particle Swarm Optimization (PSO) and fuzzy algorithm, respectively. In the practical case, Energy Storage Systems (ESSs) are indispensable parts of VSG for primary power supply to simulate the mechanical rotating energy of synchronous generators. The Hybrid Energy Storage System (HESS)-based VSG strategy is addressed in [14], but this study does not consider the rapid power transient. Hence, it is difficult to demonstrate the general effectiveness of the proposed strategy. In [15], the VSG control using the Electric Double Layer Capacitor (EDLC)-based ESS has presented to respond to the sudden load change in an islanded MG with the presence of a gas engine generator. However, the system configuration and control scheme are too simple with only a single type of generators.

In this paper, the VSG control with an additional Supercapacitor (SC)-based ESS is presented for an islanded





MG including PV system, SC, diesel generator and connected load. The studied microgrid is built using the averaged models of the converters and the mathematical equations based on the relation of the physical term in the whole system. The study concentrates on the analysis of the dynamic stability in the MG when the rapid power change occurs due to the variation of solar radiance and faults in the MG. The rest of this paper is organized as follows. Section II describes the configuration and mathematical models of the studied MG. In section III, the control strategies of VSG and the converters based on the cascaded control are presented. The simulation results in time-domain are analyzed in Section IV and the conclusion is made in Section V.

## II. SYSTEM CONFIGURATION

### A. System Description

Fig. 1 shows the configuration of the studied microgrid system. The right-hand side of Fig. 1 represents a 1.8-MW diesel synchronous generator connected to the bus of PV system through a 0.4/22-kV step-up transformer and a 50-m connection cable. The left-hand side of Fig. 1 represents a PV system and SC connected to each other through a short connection cable. In PV system, the 300-kWp PV arrays are connected to DC-link through a DC/DC boost converter equipped with the Maximum Power Point Tracking (MPPT) controller [16]. This MPPT algorithm can track the maximum power point when the solar irradiance changes and DC-link is interfaced with PV-bus via DC/AC voltage-source inverter, filters and a 0.4/22-kV step-up transformer. The connection of SC is similar to the PV system. The DC/DC bidirectional converter of SC system can transfer two-direction power to microgrid. The VSI of SC is equipped with VSG control strategy. In this configuration, a max load system of 1.2 MW is connected to the Point of Common Coupling (PCC) of the SC system and the shunt capacitor is also connected to PCC. The control strategies of this microgrid configuration will be presented in the next section of the study. The nonlinear mathematical models of the microgrid employed in this study will be described as follows.

### B. PV Array and Boost Converter Model

The PV array model can be extended from the equivalent circuit model of a photovoltaic cell in a PV module. The output current of PV array (in ampere) is expressed in the following equation [17-18]:

$$I_{PV} = N_{mp} I_{ph} - N_{mp} I_0 - \left\{ \exp \left[ \frac{q(V_{PV} + R_{es} I_{PV})}{k_{\beta} \eta_d T N_s N_{ms}} \right] - 1 \right\} - \frac{V_{PV} + R_{es} I_{PV}}{R_{ep}} \quad (1)$$

where  $V_{PV}$  is terminal voltage (V) of PV array,  $N_{mp}$  and  $N_{ms}$  are the parallel and series PV module numbers in the PV array, respectively;  $N_s$  is series-cell number in a PV module;  $R_{es}$  and  $R_{ep}$  are the series and parallel equivalent resistances ( $\Omega$ ) of the PV array, respectively;  $I_{ph}$  is the current source (A) of a PV cell;  $I_0$  is the inverting saturation current (A) of a PV cell in the standard temperature;  $q$  is the charge (C) of an electron particle;  $k_{\beta}$  is the Boltzmann constant (J/K) and  $\eta_d$  is the ideality factor of diode.

By neglecting conduction and the switching losses in power semiconductors, the averaged model of the DC/DC boost converter connected to the DC-link of the PV system is shown as Fig. 2 [20].

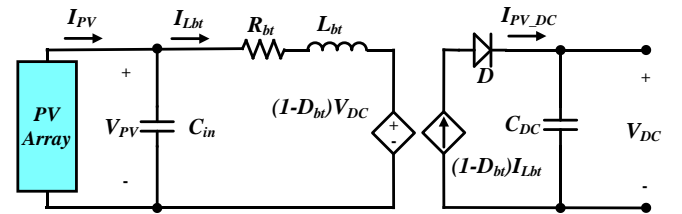


Fig. 2. The averaged model of the DC/DC boost converter of the PV system

The differential equations [18] representing for the above model are expressed by:

$$C_{in} \frac{dV_{PV}}{dt} = I_{PV} - I_{Lbt} \quad (2)$$

$$L_{bt} \frac{dI_{Lbt}}{dt} + R_{bt} I_{Lbt} = V_{PV} - (1 - D_{bt}) V_{DC} \quad (3)$$

$$I_{PV\_DC} = (1 - D_{bt}) I_{Lbt} \quad (4)$$

where  $C_{in}$  is the input capacitor of boost converter;  $R_{bt}$  and  $L_{bt}$  are resistance and reactance of the boost converter inductor;  $I_{Lbt}$  is the current flown through the inductor;  $I_{PV\_DC}$  is the out current of the boost converter;  $V_{DC}$  is the DC-link voltage and  $D_{bt}$  is the duty cycle of the boost converter.

### C. SC and Bidirectional Converter Model

In this study, the used SC model is a classic equivalent circuit model [19] expressed by the following equations:

$$C_{SC} \frac{dV_{CSC}}{dt} = -I_{SC} - \frac{V_{CSC}}{R_{pSC}} \quad (5)$$

$$V_{SC} = V_{CSC} - R_{sSC} I_{SC} \quad (6)$$

where  $C_{SC}$  is the capacitor of SC;  $R_{pSC}$  and  $R_{sSC}$  are the parallel and series resistance, respectively;  $V_{CSC}$  is the voltage of the  $C_{SC}$ -capacitor;  $V_{SC}$  is the output voltage of the SC unit;  $I_{SC}$  is the output current of the SC.

The same as the boost converter of the PV system, the averaged model of the bidirectional converter of the SC system is also built by neglecting the switching of the semiconductor as shown in Fig. 3 [20]. The schematic diagram of the bidirectional can be referred in detail in [19].

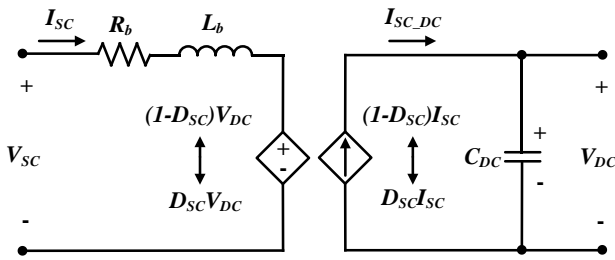


Fig. 3. The averaged model of the bidirectional converter of the SC system

The dynamic equations describing the mathematical model of the DC/DC bidirectional converter are represented following:

$$L_b \frac{dI_{SC}}{dt} + R_b I_{SC} = V_{SC} - DV_{DC} \quad (7)$$

$$I_{SC-DC} = DI_{SC} \quad (8)$$

where  $D = D_{SC}$  in the buck mode and  $D = 1 - D_{SC}$  in the boost mode;  $D_{SC}$  is the duty cycle of converter;  $R_b$  and  $L_b$  are resistance and reactance of the bidirectional converter inductor, respectively;  $I_{SC-DC}$  is the output current of converter.

### D. VSI Model and Filters

Because the switches in the semiconductors are not part of this study, it is assumed that the switching frequency of the gates of the semiconductors is much higher than the fundamental frequency of the microgrid system. Thus, the ripples of AC signals caused by switches are properly neglected, and the average model technique [21] in a  $dq$ -axis reference frame of the VSI unit is utilized for the both of the PV system and the SC system as in fig. 4. The voltage of the DC-link is kept at a constant value by regulating the energy balance at the DC-link of the VSI. As a result, the VSI terminal voltages in the  $dq$ -axis reference frame can be expressed by the DC-link voltage  $V_{dc}$  and the modulation index  $m_{dq}$ . The mathematical equations in per unit describing the behavior of the VSI are expressed in the  $dq$ -frame as follows:

$$\frac{L_f}{\omega_{base}} \frac{dI_{qVSI}}{dt} = L_f \omega I_{qVSI} - R_f I_{qVSI} + V_{cd} - V_{dPCC} \quad (9)$$

$$\frac{L_f}{\omega_{base}} \frac{dI_{qVSI}}{dt} = -L_f \omega I_{dVSI} - R_f I_{qVSI} + V_{cd} - V_{dPCC} \quad (10)$$

$$\frac{C_f}{\omega_{base}} \frac{dV_{dPCC}}{dt} = I_{dVSI} - I_{dL} + \omega C_f V_{qPCC} \quad (11)$$

$$\frac{C_f}{\omega_{base}} \frac{dV_{qPCC}}{dt} = I_{qVSI} - I_{qL} - \omega C_f V_{dPCC} \quad (12)$$

where  $R_f$ ,  $L_f$  and  $C_f$  are the resistance, inductance and capacitance of the VSI filter, respectively;  $I_{dVSI}$  and  $I_{qVSI}$  are the VSI current components in  $dq$ -frame flow through the filter;  $I_{dL}$  and  $I_{qL}$  are the load current components in  $dq$ -frame;  $\omega$  and  $\omega_{base}$  is the actual and base angular frequency of the microgrid;  $V_{cd}$ ,  $V_{cq}$  and  $V_{dPCC}$ ,  $V_{qPCC}$  are the VSI terminal voltage components and the voltages of the bus connected to the microgrid, in which  $V_{cd} = m_d V_{DC}/2$  and  $V_{cq} = m_q V_{DC}/2$ .

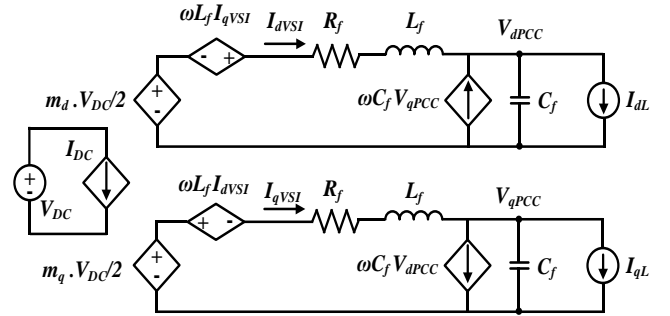


Fig. 4. The averaged model of the voltage-source inverter

### E. Model of Diesel Generator and Load

The diesel generator is modelled in three components: the prime mover, the synchronous generator and the automatic voltage regulator (AVR). The model of the electrical and mechanical part of SG using the two-axis model [22] is expressed in per-unit follows:

$$T'_{qo} \frac{dE'_d}{dt} = -E'_d + (X_q - X'_q) I_q \quad (13)$$

$$T'_{do} \frac{dE'_q}{dt} = -E'_q - (X_d - X'_d) I_d + E_{fd} \quad (14)$$

$$V_{td} = E'_d - R_s I_d + X'_q I_q \quad (15)$$

$$V_{tq} = E'_q - R_s I_q - X'_d I_d \quad (16)$$

$$2H_d \frac{d\omega_e}{dt} + D_d (\omega_e - 1) = T_{md} - (E'_d I_d + E'_q I_q) + (X'_q - X'_d) I_d I_q \quad (17)$$

$$\frac{d\theta_e}{dt} = (\omega_e - 1) \omega_{base} \quad (18)$$

where  $T'_{do}$  and  $T'_{qo}$  are the  $dq$ -components of the transient open time constants (s);  $E'_d$  and  $E'_q$  are the voltages behind transient impedance;  $I_d$  and  $I_q$  are the  $dq$ -currents through the stator winding,  $E_{fd}$  is the field voltage;  $X_d$  and  $X_q$  are the direct and quadrature axis inductances, respectively;  $X'_d$  and  $X'_q$  are the direct and quadrature axis transient inductances,

respectively;  $R_s$  is the resistance of the stator winding;  $V_{td}$  and  $V_{tq}$  are the terminal voltages of diesel generator, respectively;  $\theta_e$  is the angular of the microgrid voltage phasor (rad);  $\omega_e$  and  $\omega_{base}$  are the actual and base angular frequency of the MG, respectively;  $T_{md}$  is the mechanical torque;  $H_d$  is the inertia constant and  $D_d$  is the damping factor.

The models of the prime mover and the AVR which added the secondary loop to form a PI controller of the diesel generator utilized in this study are shown in block scheme of Fig. 5 and Fig. 6, respectively, referring in detail in [23].

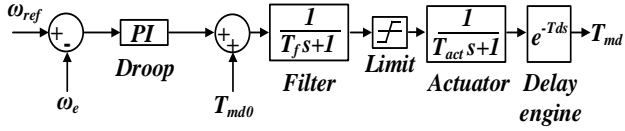


Fig. 5. The block scheme of the prime mover of the diesel generator

The well-known reduced admittance matrix is utilized here for modelling load, buses and cables. In this study, it is assumed that the impact of load, cable as a constant admittance, so the admittance matrix is used [22]. It is also considered that the load only consumes active power.

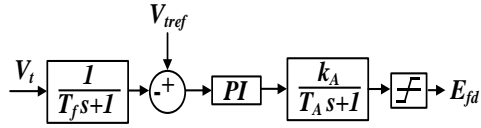


Fig. 6. The block scheme of the AVR of the diesel generator

### III. CONTROL SCHEMES OF THE STUDIED SYSTEM

In this model, the MPPT controller is controlled by the P&O algorithm [24], and the simulation results are shown in Section IV. The control block scheme of VSG is illustrated in Fig. 7 [25].

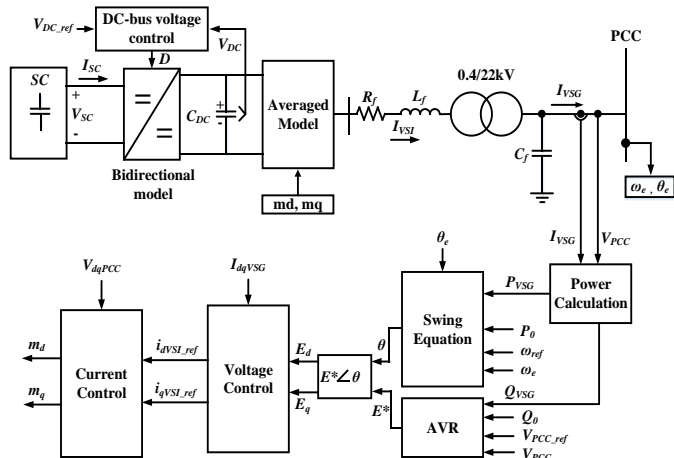


Fig. 7. Block scheme of VSG control

This scheme includes five main functions: swing equation block, AVR block, voltage control loop, current control loop and DC-link voltage control block of the bidirectional converter. As mentioned earlier, the VSG control is embedded into the control algorithm of VSI to simulate the behaviors of a traditional synchronous generator. The swing equation that is the core of VSG frequency controller in per-unit is expressed in (20) [26].

$$T_m - T_{VSG} = 2H \frac{d\omega_{VSG}}{dt} + D(\omega_{VSG} - \omega_{ref}) \quad (19)$$

where  $T_m$  is the virtual mechanical torque of VSG;  $T_{VSG}$  is the electromagnetic torque of VSG;  $\omega_{VSG}$  and  $\omega_{ref}$  are the VSG and reference angular frequency; The  $H$  and  $D$  values are the inertia constant and damping factor of VSG, respectively.

When the VSI of DGs operates in a weak grid such as microgrid, it should have ability to support the MG's frequency for the sake of stability. Hence, the droop control is applied to calculate the virtual mechanical torque. The virtual mechanical torque in pu  $T_m$  is expressed as in (20)

$$T_m = T_{m0} + \Delta T_m = T_{m0} + \frac{1}{R_d} (\omega_{ref} - \omega_e) \quad (20)$$

where  $T_{m0}$  is the initial mechanical torque of VSG;  $\omega_e$  is the angular frequency of MG and  $R_d$  is the droop constant.

However, there is a certain deviation between the angular frequency of MG and the reference angular frequency in the steady state. Thus, a secondary loop is also introduced in order to recover the MG frequency to reference value. The block scheme of regulating frequency of VSG is shown in Fig. 8.

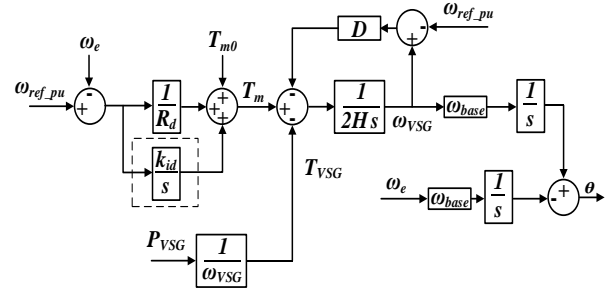


Fig. 8. Frequency control scheme of VSG

Like the diesel generator, the automatic voltage regulator (AVR) of VSG is utilized to regulate the actual VSI output voltage  $V_{PCC}$  to match the reference value of VSI output voltage  $V_{PCC\_ref}$ . The AVR output is the reference of VSI output voltage. The block scheme of AVR is shown in Fig. 9, where  $Q_0$  and  $Q_{VSG}$  are the reference reactive power and VSI output reactive power of VSG, respectively;  $D_q$  is the  $Q$ -V droop coefficient.

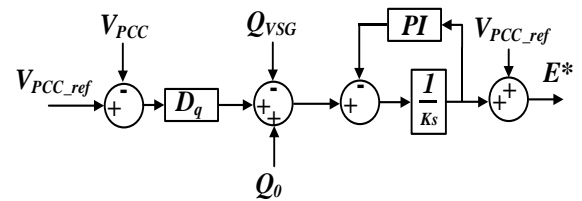


Fig. 9. Block scheme of AVR

Fig. 10 shows the block scheme of the VSI controller according to the hierarchical structure control [26] with the outer voltage control loop and the inner current control loop. The VSI controller is performed in  $dq$ -axis frame. After calculating  $E_d$  and  $E_q$  from the VSG controller,  $E_d$  and  $E_q$  are taken as reference values for the voltage control loop. The voltage control loop aims to keep stabilizing the PCC voltage, and the output reference current components in  $dq$ -frame is fed to the current decoupling controller. The current decoupling controller performs to control the current flown

through the VSI track the reference current from the voltage control. The  $d$ -axis and  $q$ -axis reference voltages generated from the hierarchical controller are used to calculate the modulation indexes  $m_d$  and  $m_q$ .

The purpose of VSG control in this study is to respond rapidly to the power transients in the MG. Thus, it is important to have an energy supply in short time for the demands of VSG-based VSI. During the operation of VSG-based VSI, the transferred power through the VSI can make the DC-link voltage deviate from the nominal value. So, the DC-link voltage controller is necessary for the VSG-based VSI to keep the DC-link voltage at the constant value. The block diagram of DC-link controller is depicted in Fig. 11 [28]. In this diagram, the reference current of the SC is generated from the voltage control loop, then the inner current control loop performs to track the reference current value of the SC and produce the duty cycle  $D_{SC}$  to send to the averaged model of the DC/DC bidirectional converter that is mentioned in Section II.

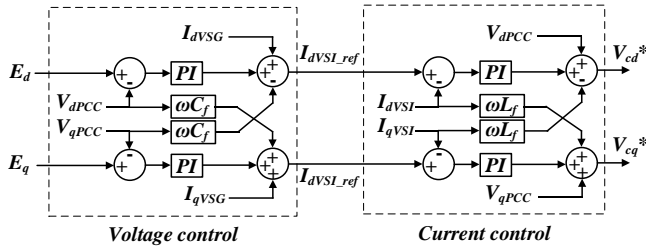


Fig. 10. Cascaded control structure of the VSI

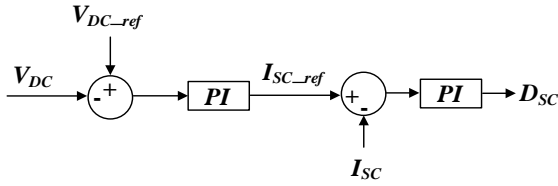


Fig. 11. DC-link voltage control

#### IV. SIMULATION RESULTS

To validate the effectiveness of the proposed VSG, the nonlinear simulations of the studied MG are established under the Matlab/Simulink environment, and the MG is constructed by utilizing the averaged mathematical models presented in Section II. Two scenarios with the change of solar irradiance and the short-circuit incidents are performed for the analysis of dynamic frequency responses of the MG. The main parameters of models are listed in Table I. This study is only concerned with frequency response, so the results of AC voltages are not given in this section.

##### A. Case 1: Dynamic Response of the MG under the Variation of Solar Irradiance

The scenario of solar irradiance is depicted in Fig. 12. The times of sudden changes in solar radiance are at 50 s, 150 s and 500 s, respectively. The time at 500 s is the toughest moment with the radiance suddenly decreases from 1000 W/m<sup>2</sup> to 200 W/m<sup>2</sup>. Fig. 13 plots the output power of the PV system using the P&O-based MPPT controller. It can be seen that the plot of the radiance and the PV output power is quite identical to each other.

Fig. 14(a) shows the dynamic frequency response of MG under the fluctuation of PV system output power. It can be

seen that the frequency deviation is greatly improved in the presence of VSG compared to the case without VSG. With VSG, the Rate of Change of Frequency (RoCoF) is slower due to the inertia of VSG and the frequency deviation from the rated value is also significantly reduced, thanks to the damping component of VSG (see the zoomed plot). As mentioned in Section II, the secondary loop of VSG and diesel generator are introduced to recover the frequency to the nominal value in steady state. The zoomed plot shows that the frequency is recovered to its rated value after 5 s. With a given grid code, the designed VSG controller satisfies the grid code by utilizing the adjustment of the inertia constant  $H$  and the damping coefficient  $D$  of VSG.

TABLE I. MAIN PARAMETERS OF SYSTEM

<b>300-kW PV array and DC/DC boost converter</b>	
1) PV module: $P_{mpp} = 305.2$ W, $V_{mpp} = 54.7$ V, $I_{mpp} = 5.58$ A, $N_s = 96$ cells, $V_{oc,n} = 64.2$ V, $I_{sc,n} = 5.96$ A, $R_s = 0.037998$ $\Omega$ , $R_p = 993.51$ $\Omega$	
2) PV array: $N_{ms} = 11$ , $N_{mp} = 90$ , $R_{es} = 4.644 \times 10^{-3}$ $\Omega$ , $R_{ep} = 121.429$ $\Omega$	
3) Boost converter: $C_{in} = 300$ $\mu$ F, $R_{bt} = 1.6667$ m $\Omega$ , $L_{bt} = 1.6667$ mH	
<b>DC-link and VSI of VSG</b>	
$V_{DC} = 1000$ V, $C_{DC} = 0.2$ F $R_f = 1.63$ m $\Omega$ , $L_f = 100$ $\mu$ H, $C_f = 2500$ $\mu$ F	
<b>SC and bidirectional converter</b>	
1) SC model: $C_{SC} = 10$ F, $R_{sSC} = 0.01$ $\Omega$ , $R_{pSC} = 10^4$ $\Omega$ , 2) Bidirectional dc/dc converter: $R_b = 1$ m $\Omega$ , $L_b = 2$ mH	
<b>VSG</b>	
$R_d = 0.05$ , $H = 3$ s, $D = 20$ , $k_{id} = 20$ , $\omega_{ref} = 1$ pu, $\omega_{base} = 314.16$ rad/s $V_{PCC\_ref} = 1$ pu, $K = 12$ , $D_q = 166.7$	
<b>1.8-MW diesel generator</b>	
$X_d = 1.22$ , $X'_d = 0.174$ , $X_q = 1.16$ , $X'_q = 0.25$ , $T'_{do} = 5.2$ , $T'_{go} = 0.5$ , $H_d = 0.495$ , $D_d = 0.0003$	

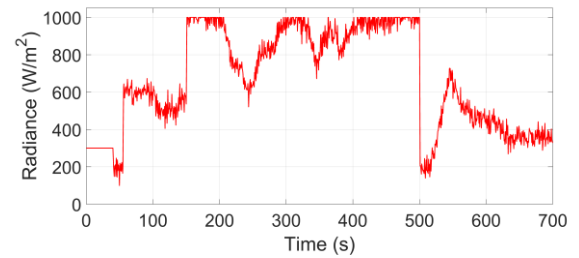


Fig. 12. Variation of solar radiance

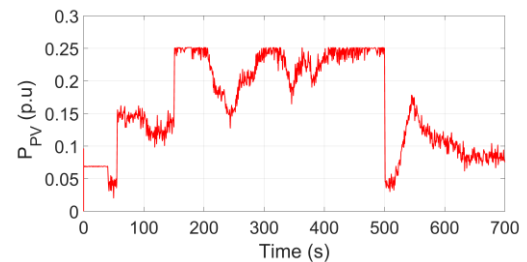


Fig. 13. Fluctuation of output PV power

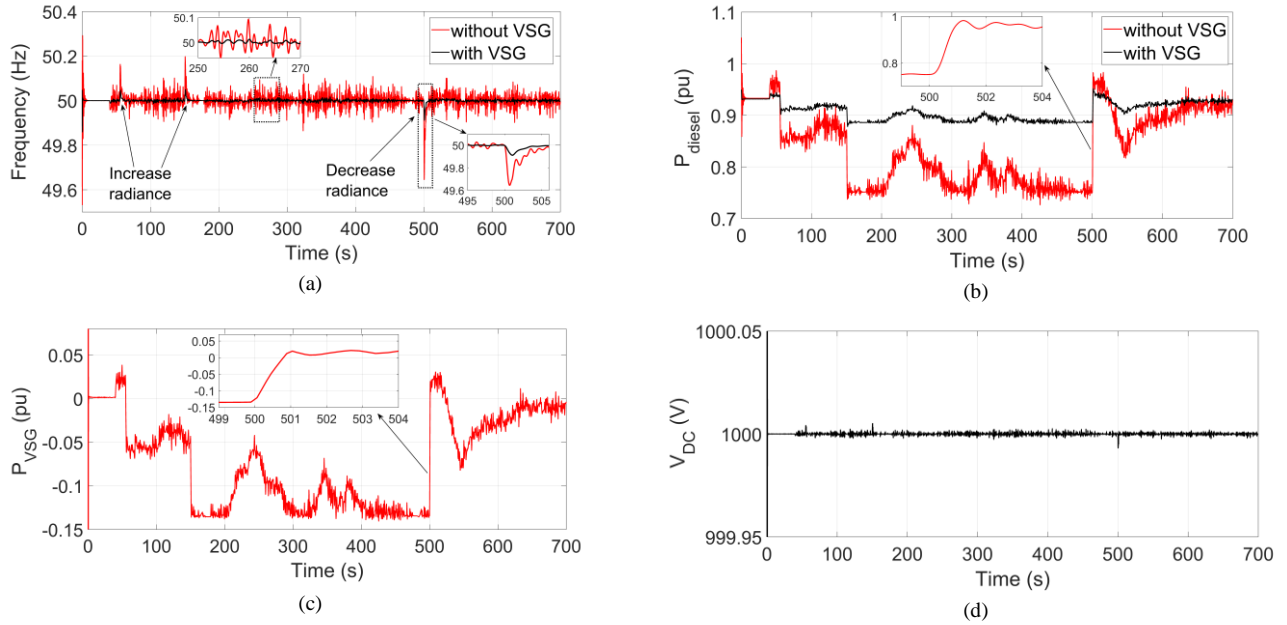


Fig. 14. Dynamic response of the studied MG under the variation of solar radiance

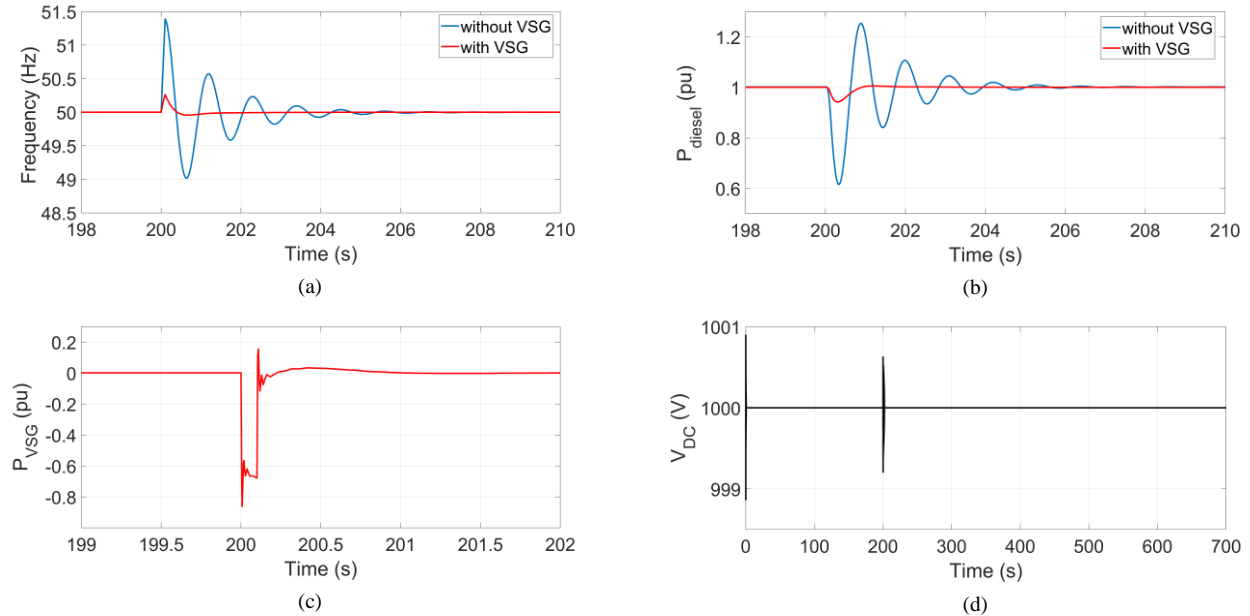


Fig. 15. Dynamic response of the studied MG under the circuit short fault

Fig. 14(b) and 14(c) show the power responses of the VSI output power of VSG and the mechanical power of diesel generator. The mechanical power waveform of the diesel generator fluctuates greatly compared to the case without VSG control, which causes instability for the diesel's rotor. A large engine speed deviation occurs, and the engine speed is beyond the permissible range. In the zoomed plot of Fig. 14 (b), the diesel generator cannot respond immediately at the time the transient occurs due to the engine delay and the actuator mechanisms (can be seen at the time of 500 s). Hence, the frequency deviation of MG is very large. Fig. 14 (c) shows that the variation in the diesel generator is greatly improved by using of VSG control, and the response time of the VSG is fast enough to reduce the instability in the MG as well as the diesel generator. Fig. 14(d) depicts the DC-link voltage waveform. It can be clearly observed that the DC-link voltage has some slight ripples under the fluctuation of PV. However,

it is negligible thanks to DC-link controller and the quick response ability of SC.

#### B. Case 2: Dynamic Response of the MG under the Short Circuit Fault

In this case, the transient response of MG under the three-phase short-circuit fault is evaluated when the fault is begun at  $t = 200$  s and lasted for 0.1 s, applied to the bus of loads. The load at this time is reduced from 1 pu to 0.2 pu. Moreover, the duration of the short circuit fault is very small, so the PV power fluctuations are not considered in this case.

Fig. 15(a) depicts the frequency of MG when the fault occurs at  $t = 200$  s. Without VSG, the overshoot of MG frequency is quite large and there is a low-frequency oscillation after the fault is cleared. The low-frequency oscillation is damped after about 6 s. This makes the MG unstable and leads to a decrease in the power quality of the



MG. On the contrary, the frequency response is greatly improved with the overshoot reduced significantly and without the low-frequency oscillation in the waveform of the frequency in the case of VSG control. Similar to the frequency, the mechanical power response of the diesel generator also has a large oscillation and is suppressed after about 6 s in the absence of VSG control and is improved with the appearance of VSG as in Fig. 15(b). The output power waveform of VSG is depicted in Fig. 15(c). Thanks to the inertia, the output power of VSG can respond quickly to sudden transient due to the short circuit fault. The response time is in the range of milliseconds thanks to the SC energy supply. Fig. 15(d) indicates DC-link voltage waveform. There is a sudden change of DC-link voltage at the time of the fault, however it is insignificant.

## V. CONCLUSION

In order to improve the dynamic performance of the islanded MG, using VSG control is extremely important. In this paper, VSG is introduced into a studied MG including a PV power generation, diesel generator and the connection-load. The VSG control is built from the dynamic characteristics of a traditional synchronous generator, integrated with the cascaded control structure of the voltage-source inverter to transfer the inertia power to the MG, thus enhancing the power quality of MG during the transients. Supercapacitor-based the energy storage system is also introduced in this study as a fast response energy supply, and it is considered as the primary energy of the VSG. The mathematical models and simulation results in time-domain verify the dynamic stability of the MG with the presence of the VSG control.

## ACKNOWLEDGMENT

This research was funded by the Ministry of Education and Training under project number B2019-DNA-11.

## REFERENCES

- [1] R. C. Dugan and T. E. McDermott, "Distributed generation," *IEEE Ind. Appl. Mag.*, vol. 8, no. 2, pp. 19–25, Mar./Apr. 2002.
- [2] R. H. Lasseter, "MicroGrids," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, pp. 305–308, 2002.
- [3] H. Bevrani, T. Ise, and Y. Miura, "Virtual synchronous generators: A survey and new perspectives," *Electrical Power & Energy Systems*, vol. 54, pp. 244–254, January 2014.
- [4] T. Kerdphol, F. S. Rahman, and Y. Mitani, "Virtual Inertia Control Application to Enhance Frequency Stability of Interconnected Power Systems with High Renewable Energy Penetration," *Energies* 2018, vol. 11, no. 4, pp. 1–16, April 2018.
- [5] R. Shi, X. Zhang, C. Hu, H. Xu, J. Gu, W. Cao, "Self-tuning virtual synchronous generator control for improving frequency stability in autonomous photovoltaic-diesel microgrids," *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 3, pp. 482–494, December 2017.
- [6] J. Liu, Y. Miura, and T. Ise, "Comparison of Dynamic Characteristics Between Virtual Synchronous Generator and Droop Control in Inverter-Based Distributed Generators," *IEEE Trans on Power Electronics*, vol. 31, no. 5, pp. 3600–3611, May 2016.
- [7] M. A. Torres, L. A. C. Lopes, L. A. Moran, and J. R. Espinoza, "Self-Tuning Virtual Synchronous Machine: A Control Strategy for Energy Storage Systems to Support Dynamic Frequency Control," *IEEE Trans on Energy Conversion*, vol. 29, no. 4, pp. 833–840, Dec. 2014.
- [8] J. Alipoor, Y. Miura, and T. Ise, "Power System Stabilization Using Virtual Synchronous Generator With Alternating Moment of Inertia," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 3, no. 2, pp. 451–458, June 2015.
- [9] N. Soni, S. Doolla, and M. C. Chandorkar, "Improvement of Transient Response in Microgrids Using Virtual Inertia," *IEEE Trans on Power Delivery*, vol. 28, no. 3, pp. 1830–1838, July 2013.
- [10] A. B. Chowdhury, X. Liang, and H. Zhang, "Fuzzy-Secondary-Controller-Based Virtual Synchronous Generator Control Scheme for Interfacing Inverters of Renewable Distributed Generation in Microgrids," *IEEE Trans on Industry Applications*, vol. 54, no. 2, pp. 1047–1061, March–April 2018.
- [11] J. Fang, H. Li, Y. Tang, and F. Blaabjerg, "Distributed Power System Virtual Inertia Implemented by Grid-Connected Power Converters," *IEEE Trans on Power Electronics*, vol. 33, no. 10, pp. 8488–8499, Oct. 2018.
- [12] B. Rathore, S. Chakrabarti, and S. Anand, "Frequency response improvement in microgrid using optimized VSG control," *2016 National Power Systems Conference*, December 2016.
- [13] Y. Hu, W. Wei, Y. Peng, and J. Lei, "Fuzzy virtual inertia control for virtual synchronous generator," *2016 35<sup>th</sup> Chinese Control Conference*, pp. 8523–8527, July 2016.
- [14] J. Fang, X. Li, Y. Tang, H. Li, "Power management of virtual synchronous generators through using hybrid energy storage systems," in *Proceedings of the Applied Power Electronics Conference and Exposition*, pp. 1407–1411, March 2018.
- [15] H. S. Hlaing, J. Liu, Y. Miura, H. Bevrani, and T. Ise, "Enhanced Performance of a Stand-Alone Gas-Engine Generator Using Virtual Synchronous Generator and Energy Storage System," *IEEE Access*, vol. 7, pp. 176960–176970, Dec. 2019.
- [16] M. Hlaaili, H. Mechergui, "Comparison of Different MPPT Algorithms with a Proposed One Using a Power Estimator for Grid Connected PV Systems," *International Journal of Photoenergy*, pp. 1–10, Jun. 2016.
- [17] V. T. Nguyen, B. N. Nguyen, H. H. Nguyen, K. H. Le, M. Q. Duong, H. L. Le, "A Proposal for an MPPT Algorithm Based on the Fluctuations of the PV Output Power, Output Voltage, and Control Duty Cycle for Improving the Performance of PV Systems in Microgrid," *Energies*, vol. 13, no. 17, 2020.
- [18] B. N. Nguyen, V. T. Nguyen, M. Q. Duong, K. H. Le, H. H. Nguyen, A. T. Doan, "Propose a MPPT algorithm based on Thevenin equivalent circuit for improving photovoltaic system operation," *Frontier in Energy Research*, Vol. 8, Feb. 2020.
- [19] L. Wang, Q.-S. Vo, and A. V. Prokhorov, "Dynamic stability analysis of a hybrid wave and photovoltaic power generation system integrated into a distribution power grid," *IEEE Trans. Sustainable Energy*, vol. 8, no. 1, pp. 404–413, Jan. 2017.
- [20] E. V. Dijk, H. J. N. Spruijt, D. M. O'Sullivan, and J. B. Klaassens, "PWM-switch modeling of DC-DC converters," *IEEE Trans. Power Electronics*, vol. 10, no. 6, pp. 659–665, Nov. 1995.
- [21] J. Sun and H. Grotstollen, "Averaged modelling of switching power converters: Reformulation and theoretical basis," in *Proc. IEEE PESC*, 1992, pp. 1166–1172.
- [22] P. Kundur, *Power System Stability and Control*. New York: McGraw-Hill, 1994.
- [23] A. Cuculić, J. Čelić, R. Prenc, "Marine Diesel-generator Model for Voltage and Frequency Variation Analysis During Fault Scenarios," *Pomorski zbornik*, vol. 51, no. 1, pp. 11–24, 2016.
- [24] D. Sera, L. Mathe, T. Kerekes, S. V. Spataru, and R. Teodorescu, "On the Perturb-and-Observe and Incremental Conductance MPPT Methods for PV Systems," *IEEE Journal of Photovoltaics*, vol. 3, no. 3, pp. 1070–1078, July 2013.
- [25] J. Liu, Y. Miura, H. Bevrani, T. Ise, "Enhanced Virtual Synchronous Generator Control for Parallel Inverters in Microgrids," *IEEE Trans on Smart Grid*, vol. 8, no. 5, pp. 2268–2277, Sep. 2017.
- [26] G. Yao, Z. Lu, Y. Wang, M. Benbouzid, and L. Moreau, "A Virtual Synchronous Generator Based Hierarchical Control Scheme of Distributed Generation Systems," *Energies* 2017, vol. 10, no. 12, pp. 1–23, Dec. 2017.
- [27] R. I. A. Yazdani, *Voltage-Sourced Converters in Power Systems: modeling, control, and applications*, John Wiley & Sons, 2010.
- [28] B. Dong, Y. Li, and Z. Zheng, "Control strategies of DC-link voltage in islanded operation of microgrid," *4<sup>th</sup> International Conference Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*, pp. 1671–1674, July 2011.

# 3D Numerical Simulation Study of a Pre-Heater Used in Solid Oxide Fuel Cell Technology

XuanVien Nguyen\*

Department of Thermal Engineering,  
Renewable Energy Research Center  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
viennx@hcmute.edu.vn

TrangDoanh Nguyen

Department of Thermal Engineering  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
trangdoanhnguyen@gmail.com

AnQuoc Hoang

Department of Thermal Engineering  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
hanquoc@hcmute.edu.vn

MinhHung Doan

Department of Thermal Engineering  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
hungdm@hcmute.edu.vn

ThiNhung Tran

Department of Chemical Technology  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
nhungtt@hcmute.edu.vn

**Abstract**—Implementing a higher operating temperature for a solid oxide fuel cell improves its performance. Previous studies have found that cell performance at a higher inlet temperature is greater than cell performance at a lower inlet temperature. The purpose of this work is to investigate the gas-to-gas heat transfer performance of fuel and air pre-heaters and to evaluate its impact on solid oxide fuel cell performance. This pre-heater is expected to be inexpensive to install and effective in heat transfer for solid oxide fuel cell application. This type of heat exchanger has been previously investigated for other applications, but not for solid oxide fuel cell technology. In this paper, heat transfer performance within the pre-heater is numerically solved using three-dimensional (3D) Ansys computational fluid dynamics for various temperatures and different pre-heater configurations for optimal design. The results show that the enhancement of the pre-heater's heat transfer surface area increases heat transfer inside the unit, thereby increasing the temperature at the pre-heater outlet, which is used to supply the solid oxide fuel cell system.

**Keywords**—fuel pre-heater, air pre-heater, convective heat transfer coefficient, computational fluid dynamics (CFD), heat transfer

## I. INTRODUCTION

Heat exchanger designers have always attempted to enhance heat transfer to reduce thermal equipment sizes and costs. Heat exchangers have widely been exploited so that energy loss can be avoided. To develop an optimal heat exchanger, many manufacturers are attempting to introduce a type of exchanger that is compact and has a high overall heat transfer coefficient. The development of high temperature fuel cell technology, in particular, solid oxide fuel cells (SOFCs), is involved in using heat exchangers as pre-heaters [1,2]. Previous studies have found that a higher operating temperature for SOFCs improves performance. Thus, knowing the temperature distribution within the SOFC is important to its operation. The effects of the inlet fuel and oxidant temperature on the maximum and average current densities are due to electrochemical reactions of carbon monoxide and hydrogen. The results showed that the total average current density at a higher inlet temperature is greater than the total average current density at a lower inlet temperature [3-5]. In this study, we developed a coupled three-dimensional (3D) thermo-fluid/thermo-mechanical modelling approach of a plate type air pre-heater used in solid oxide fuel

cells. The results show reduced prototype costs and product development time in the design and optimization of an efficient air pre-heater [6]. We also developed a plate heat exchanger in a bipolar plate used in a proton exchange membrane fuel cell (PEMFC), which operates at 80–90 °C and concentrates on size reduction [7]. A previous study investigated the effects of heat exchangers on the performance of SOFC, molten carbonate fuel cell (MCFC)/gas turbine hybrid, and proton exchange membrane fuel cell (PEMFC) systems. In a PEMFC system, it was demonstrated that an external reformation of natural gas and gasoline into hydrogen reduces carbon monoxide content to <10 ppm, which requires a number of reformers and heat exchangers (individual or combined). It was found that heat exchangers in MCFC and SOFC hybrid systems greatly increased the cell performance in terms of efficiency [8].

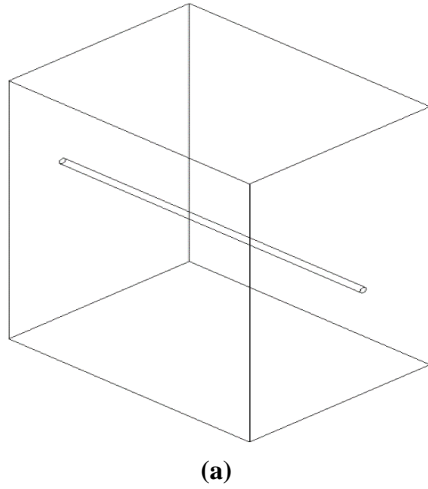
In this paper, we numerically solve heat transfer performance within the pre-heater using 3D Ansys computational fluid dynamics for various temperatures. The effects of different velocities of 0.5, 1.0, 1.5, and 2.0 m.s<sup>-1</sup> of air inlet are considered in this paper. The outlet temperature in two cases (with and without a pre-heater) was also investigated and compared.

## II. MODELING OF HEAT EXCHANGER

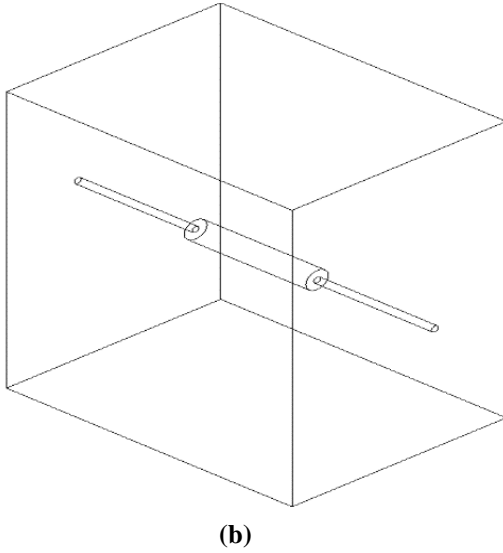
### A. Physical Model

In this study, two models were built to compare temperature outlets in the two cases (with and without the pre-heater). Figure 1a shows the geometry without the pre-heater, i.e., a pipe with a diameter of 10 mm and a length of 170 mm. The pre-heater consisted of the main cylindrical tube with a diameter of 32 mm and a length of 170 mm (as shown in Fig.1b). All of remaining characteristics were the same in cases without the pre-heater. The device was installed in front of the fuel inlet of the SOFC system to improve heat transfer to increase the temperature. It was left at a high temperature. Both models were placed in the same furnace with the following temperatures: 873 K, 973 K, and 1123 K. Heat transfer took place between the outer and inner surfaces. Then, these two parts were imported into Ansys, where the insert domain was replaced by a Boolean and cut to an adiabatic wall, thus minimizing the computational domain. Figure 1 shows the schematics of the computational model and

discretized domain. The physical and thermal properties of air at simulated conditions are shown in Table 1.



(a)



(b)

Fig. 1. Geometry and dimension of models: a) without the pre-heater and b) with the pre-heater.

TABLE 1. PHYSICAL PROPERTIES OF AIR UNDER THE SIMULATION CONDITIONS.

Properties	Value
Thermal conductivity, $k$ (W/m.K)	0.0242
Specific heat, $C_p$ (J/(kg .K)	1006.43
Air density, $\rho$ (kg/m <sup>3</sup> )	1.225
Dynamic viscosity, $\mu$ (N.s/m <sup>2</sup> )	$1.7894 \times 10^{-5}$

### B. Mathematic Equations

Computational fluid dynamics (CFD) is a technique based on numerical methods aimed at analyzing the movement of fluids, heat transfer, mass transfer, and chemical reactions. The present study seeks to examine the fluid flow and heat transfer in a heat-exchanger filled with coiled wire inserts. Therefore, the governing equations include mass conservation (continuity), momentum conservation, and energy

conservation equations: Mass conservation (continuity)

$$\text{equation: } \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = 0 \quad (1)$$

Momentum conservation equation:

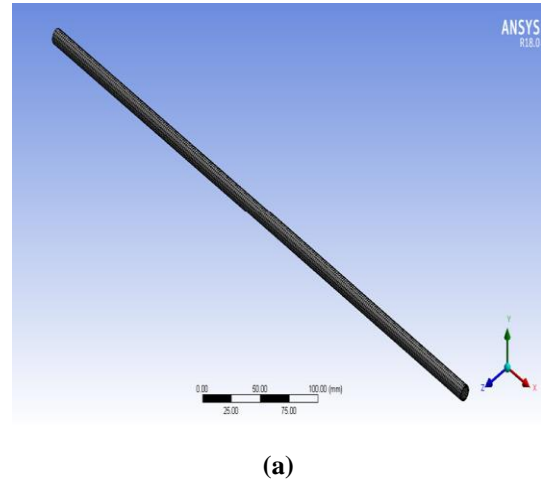
$$\frac{\partial}{\partial t} (\rho \vec{v}) + \nabla \cdot (\rho \vec{v} \vec{v}) = -\nabla p + \rho \vec{g} + \vec{f} \quad (2)$$

Energy conservation equation:

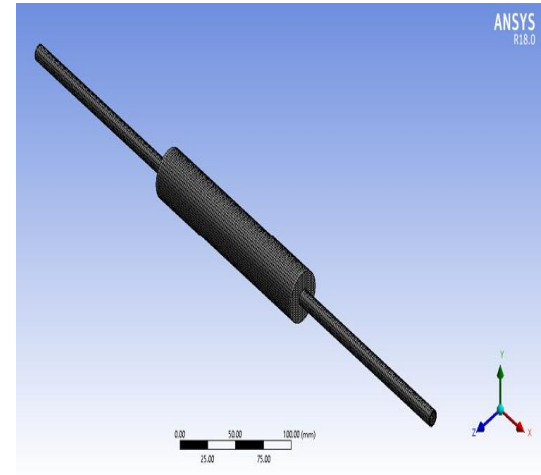
$$\frac{\partial}{\partial t} [\rho e] + \nabla \cdot [(\rho e + p) \vec{v}] = \nabla \cdot \left( k_{eff} \nabla T - \sum_j h_j \vec{J}_j + (\vec{\tau}_{eff} \cdot \vec{v}) \right) \quad (3)$$

where  $t$  is the time,  $\rho$  is the mass density,  $\vec{v}$  stands for the flow velocity,  $p$  is the fluid pressure,  $e$  is the inside energy per unit mass,  $\vec{f}$  shows the volumetric force,  $k_{eff}$  is the thermal conductivity,  $h_j$  is the enthalpy of species  $j$ ,  $\vec{J}_j$  represents the diffusion flux of species  $j$ , and finally,  $\vec{\tau}_{eff}$  exhibits the stress tensor.

### C. Mesh Generation



(a)



(b)

Fig. 2. Computational domain and mesh: a) without the pre-heater and b) with the pre-heater.

One of the most important parts of the simulation process was selecting a suitable meshing process to calculate the fundamental equations that govern the heat exchanger operation. Ansys Fluent software was used to mesh the models and simulate the system, respectively. Selecting proper meshing can contribute to suitable convergence in solving the equations, while improper meshing can lead to instability and calculation divergence. Figure 2 shows the structured hexahedral mesh algorithm that was applied throughout the whole pipe and pre-heater. This type of mesh produced more reliable results due to its logical aspect ratios and skewness. Furthermore, as this type of mesh was less diffusive than the other types, it protected the results of simulation from any inaccuracy. The model was then exported for meshing, which was a process where the model was divided into a finite number of smaller elements, with 500000.

### III. RESULTS AND DISCUSSION

#### A. The Temperature Profile of the Pipe without the Pre-heater

The results from the predicted computational fluid dynamic enabled a detailed view of temperature distribution in the pipe without the pre-heater. High temperature regions were determined, indicating regions susceptible to thermally induced stress within the component. Figure 3 illustrates how

the temperature contour plots simulated the pipe without the pre-heater at velocities of 0.5, 1.0, 1.5, and 2.0 m.s<sup>-1</sup>, respectively. As shown in Figure 3a, the outlet air temperature at a velocity of 0.5 m.s<sup>-1</sup> was the highest. The numerically calculated results revealed that smaller velocities resulted in more heat exchange, which was expected as the amount of heat transferred was greater. This was visible at higher temperatures in the pipe's outlet. The pipe's outlet temperature at velocities of 0.5, 1.0, 1.5, and 2.0 m.s<sup>-1</sup> was 865.9, 833, 799.4, and 774.7 K, respectively.

#### B. The Temperature Profile of the Pipe with the Pre-heater

Figure 4 shows the simulation results for the pipe with the pre-heater. It had a diameter of 32 mm and was 170 mm in length. For the investigation, the previously used velocities—0.5, 1.0, 1.5, and 2.0 at 873 K—were considered. The results showed how velocity effected the pre-heater's thermomechanically-induced stress behavior. As shown in Figure 4a, the air's outlet temperature had the highest velocity at 0.5 m.s<sup>-1</sup>. This was similar to the air's outlet temperature in the pipe without the pre-heater. The temperature difference between the air inlet and outlet decreased as the velocity of air at the inlet increased from 0.5 to 2.0 m.s<sup>-1</sup> due to an increase in the mass flow rate of air. The outlet temperatures for the pipe with the pre-heater at velocities of 0.5, 1.0, 1.5, and 2.0 m.s<sup>-1</sup> were 868, 847, 824.9, and 807.4 K, respectively.

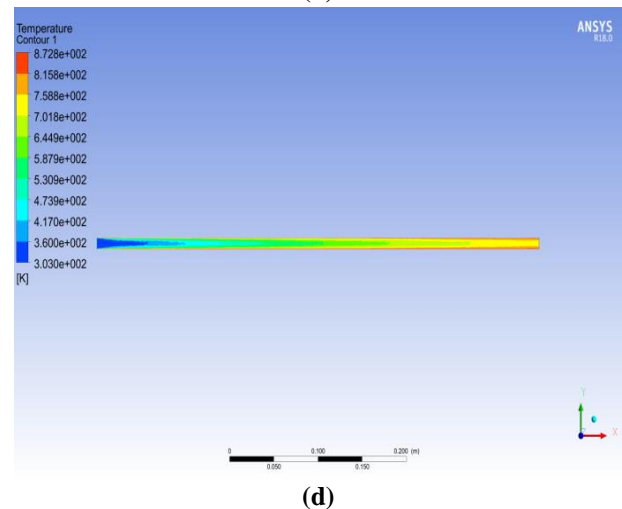
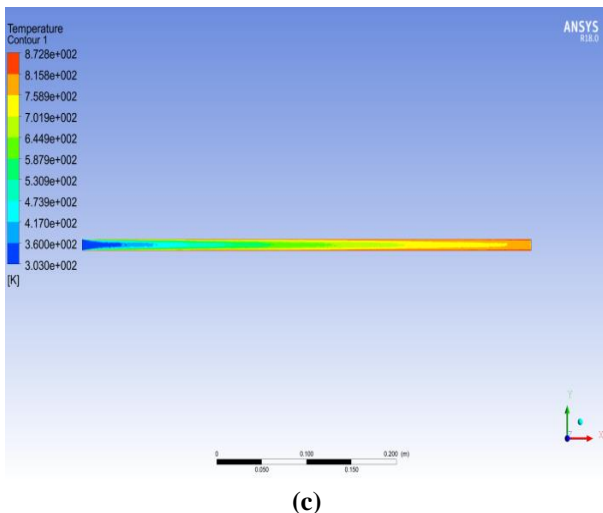
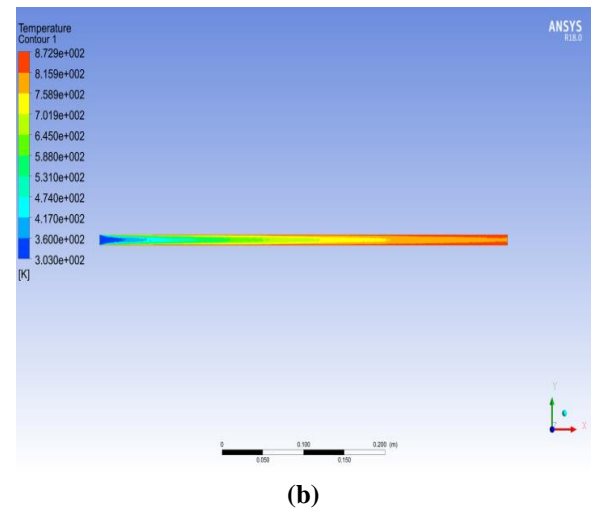
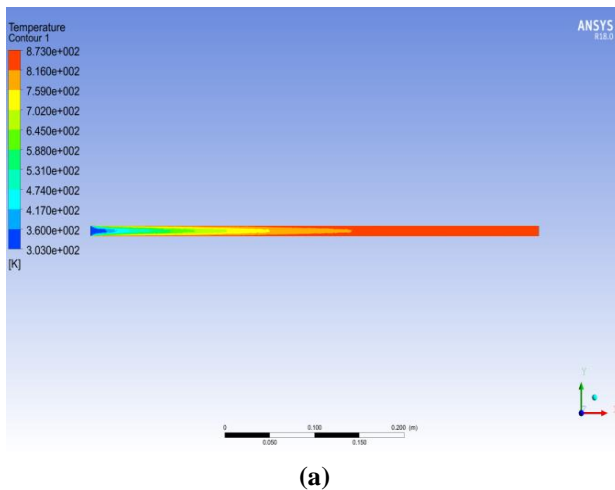


Fig. 3. Distribution of the temperature of the air outlet without the pre-heater at 873 K: a) 0.5 m/s; b) 1.0 m/s; c) 1.5 m/s; and d) 2.0 m/s.

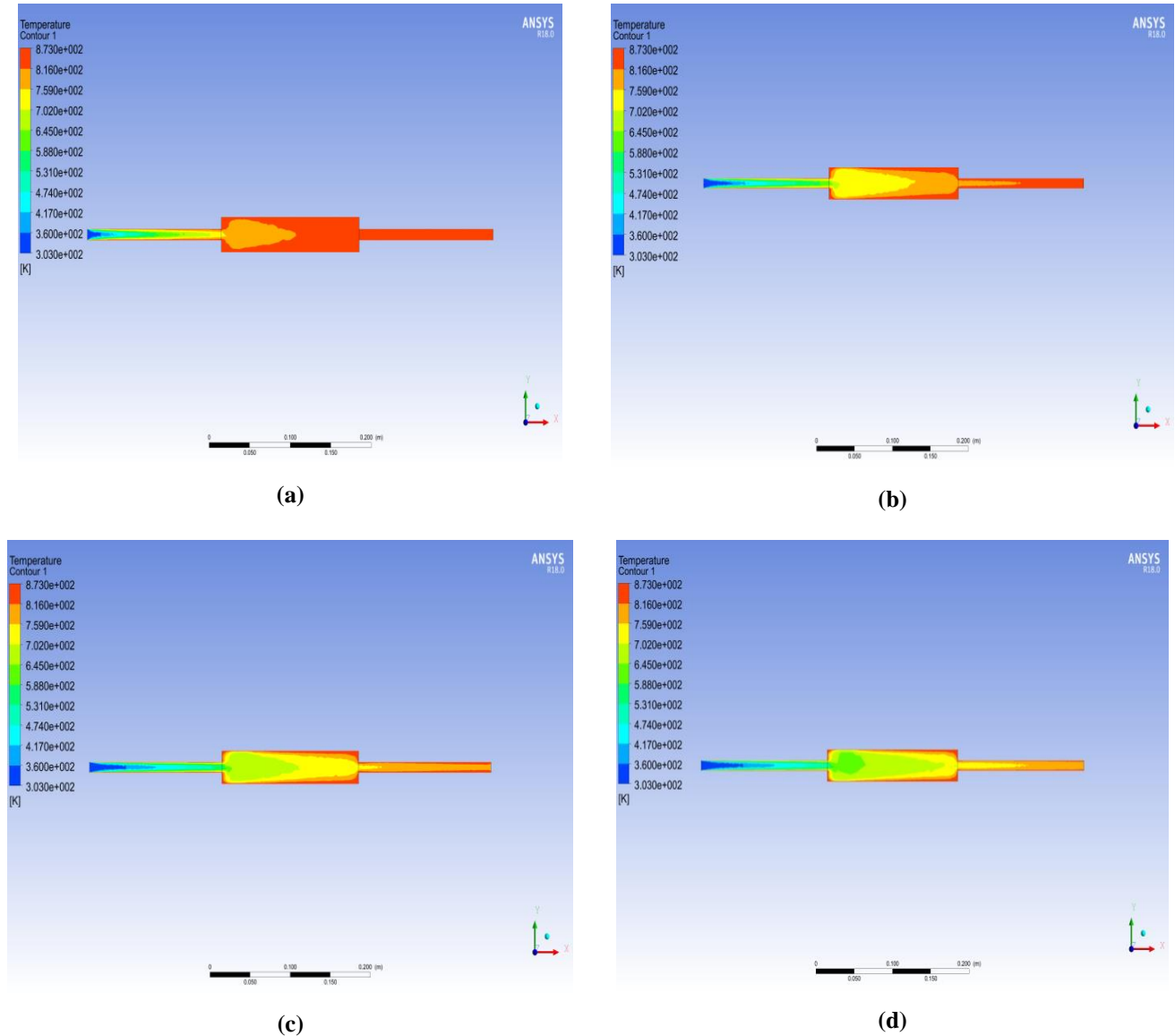


Fig. 4. Distribution of the temperature of the outlet of air with the pre-heater at 873 K: a) 0.5 m/s; b) 1.0 m/s; c) 1.5 m/s; and d) 2.0 m/s.

We implemented a comparison between the distribution of temperature along the length of the air's outlet with and without the pre-heater. Figure 5 shows the temperature difference between the pipe with and without the pre-heater. A comparison was made at an operating temperature of 873K using various velocities at the air inlet (0.5, 1.0, 1.5, and 2.0 m.s<sup>-1</sup>). At a velocity of 0.5 m.s<sup>-1</sup>, the air outlet temperature of the pipe with a pre-heater was 868 K, which was higher than the pipe without the pre-heater at 865.9 K (as shown in Fig. 5a). The temperature difference between the air inlet and outlet decreased as the velocity of air at the inlet increased from 0.5 to 2.0 m.s<sup>-1</sup> (as shown in Fig. 5b–d). The temperature

differences between cases with and without the pre-heater increased with an increasing velocity. As shown in Figure 5, from 0 to 16 cm, the temperature of the two cases (with and without pre-heater) was similar. However, from 16 to 50 cm, the temperature differences between the two cases increased. This occurred because the air moved in the pipe before entering the pre-heater (from positions of 16 to 50 cm). From 16 to 50 cm, air entered through the pre-heater and the heat exchange process occurred. The model with pre-heater assembly enhances heat conversion and increases the inlet temperature.



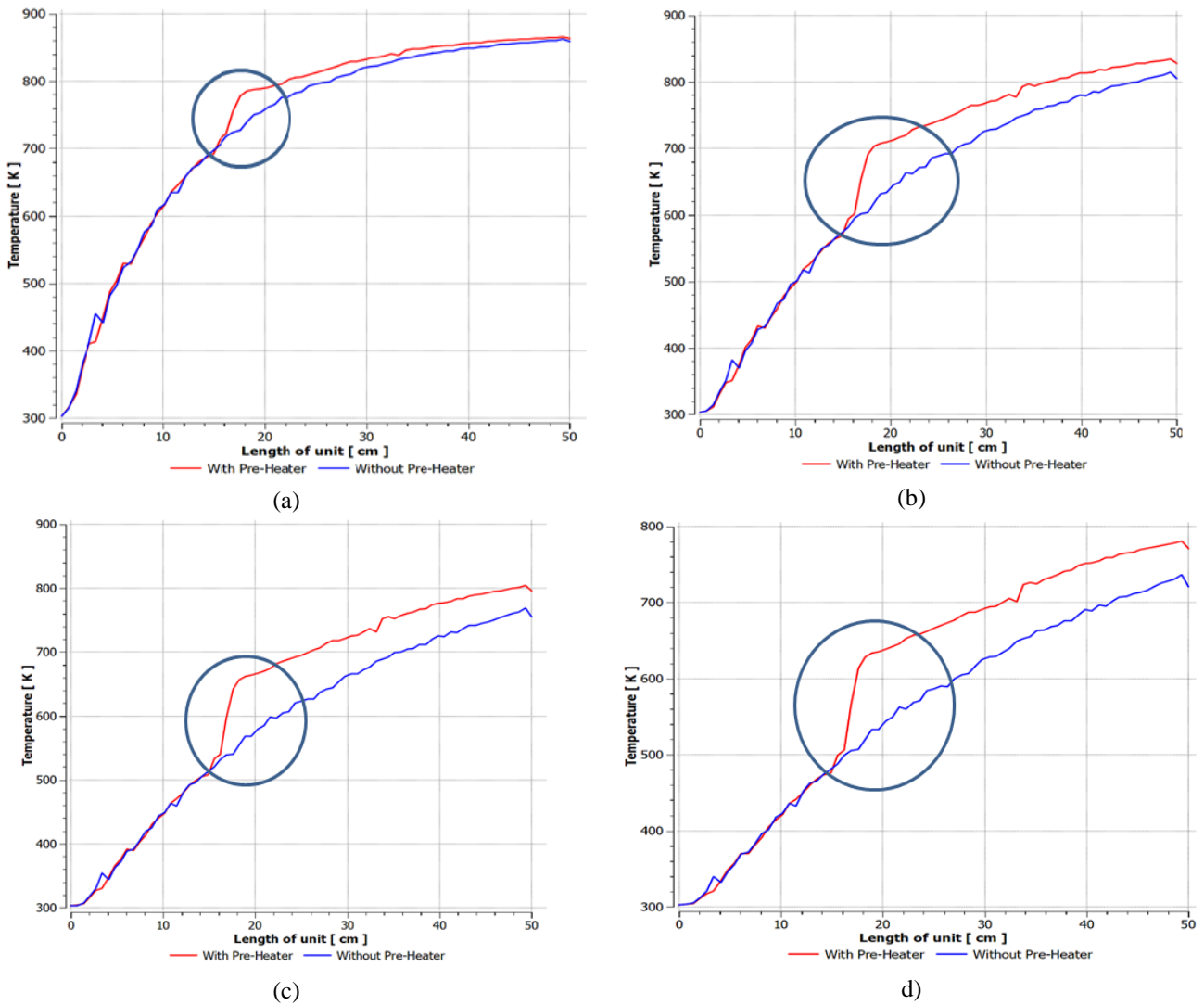


Fig. 5. The comparison of distribution temperature on the length of the outlet of air without and with the pre-heater at 873 K: a) 0.5 m/s; b) 1.0 m/s; c) 1.5 m/s; and d) 2.0 m/s.

#### IV. CONCLUSIONS

In this study, a numerical simulation study was implemented for application based on SOFC computational fluid dynamics. The outlet temperature in two cases (with and without a pre-heater) was investigated and compared. The higher operating temperatures of SOFCs resulted in improved cell performance. The system temperature, caused by the furnace, must be maintained at a sufficiently high level to ensure proper operation of the SOFC. The results indicated that the temperature with the pre-heater was higher than without the pre-heater. The pre-heater was successful at enhancing heat transfer to improve the air's inlet temperature. The results of these simulation models imply that the air's outlet temperature with the pre-heater assembly was higher than the case without the pre-heater. This result, in turn, signifies that the former system performs better than the latter, which is in agreement with the aforementioned results.

#### ACKNOWLEDGMENTS

This research was funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.03–2018.332.

#### REFERENCES

- [1] He Y, Zheng H, Zhang Q, Tang Y, Zhang J. Analysis of Enhanced Heat Transfer Performance Through Duct with Constant Wall Temperature. *J. of Thermophysics and Heat Transfer* 26 (2012) 472–475.
- [2] Liu S, Sakr M. A comprehensive review on passive heat transfer enhancements in pipe exchangers. *Renewable and Sustainable Energy Reviews* 19 (2013) 64–81.
- [3] E. Vakouftsi, G.E. Marnellos, C. Athanasiou, F. Coutelieris, CFD modeling of a biogas fuelled SOFC, *Solid State Ionics* 192 (2011) 458–463.
- [4] Joonguen Park, Peiwen Li, Joongmyeon Bae, Analysis of chemical, electro-chemical reactions and thermofluid flow in methane-fed internal reforming SOFCs: Part II: temperature effect, *Int. J. Hydrogen Energy* 37 (2012) 8532–8555.
- [5] Haddad Djamel, Abdenebi Hafsia, Zitouni Bariza, Ben Moussa Hocine, Oulmi Kafie, Thermal field in SOFC fed by hydrogen: inlet gases temperature effect, *Int. J. Hydrogen Energy* 38 (2013) 8575–8583.
- [6] M. Peksen, Ro. Peters, L. Blum, D. Stolten. 3D coupled CFD/FEM modelling and experimental validation of a planar type air pre-heater used in SOFC technology. *Int. J. Hydrogen Energy* 36 (2011) 6851–6861.
- [7] Lasbet Y, Auvity B, Castelain C, Peerhossaini H. A chaotic heat-exchanger for PEMFC cooling applications. *J. of Power Sources* 156 (2006) 114–118.
- [8] L. Magistri, A. Traverso, A.F. Massardo, R.K. Shah, Heat exchangers for fuel cell and hybrid system applications, *J. Fuel Cell Sci. Technol.* 3 (2006) 111–118.

# Deep Learning Based Semantic Segmentation for Nighttime Image

Hien T.T Bui

University of Science and  
Technology – The University  
of Danang  
Danang, Vietnam  
thuhienak04@gmail.com

Duy H. Le

University of Technology and  
Education – The University of  
Danang  
Danang, Vietnam  
lhduy@ute.udn.vn

Thu T.A Nguyen

University of Science and  
Technology - The University  
of Danang  
Danang, Vietnam  
ntathu@dut.udn.vn

Tuan V. Pham

University of Science and  
Technology - The University  
of Danang  
Danang, Vietnam  
pvtuan@dut.udn.vn

**Abstract**— Semantic segmentation of nighttime images has become an interesting research topic recently. In this work, we focus on semantic object recognition for nighttime driving scenes. The paper proposes a method to adapt the semantic models trained on daytime scenes to nighttime scenes through twilight time. In this process, the Pyramid Scene Parsing Network (PSPNet) model is suggested to provide an advanced framework for pixel prediction. The goal of the method is to reduce the cost of human annotation for nighttime scenes by transferring knowledge from typical daytime illumination conditions. Our model is trained and tested on the Cityscape dataset which is recorded in street scenes and intended for assessing the performance of vision algorithms for major tasks of semantic urban scene understanding. The proposed PSPNet model yields a mIoU record of 44.9% on nighttime driving scenes. Our experiments show that the proposed method is effective for knowledge transfer from daytime scenes to nighttime scenes without using additional human annotation. Further analysis on the proposed method has been presented in this study.

**Keywords**—semantic segmentation, driving scenes, twilight time, Pyramid Scene Parsing Network (PSPNet), knowledge transfer, deep learning.

## I. INTRODUCTION

Semantic segmentation is very important in self-driving cars which become more and more common. And safer autonomous driving at night is a necessary work in order to alleviate vehicle accidents in the dark. This is one of the reasons why people have to develop this field more especially its accuracy. One of the big reasons that automated cars have not gone mainstream yet is because it cannot deal well with nighttime and adverse weather conditions. Camera sensors can become untrustworthy at nighttime, in foggy weather, and in wet weather. Thus, computer vision systems have to function well also under these adverse conditions. In this work, we focus on semantic object recognition for nighttime driving scenes.

Semantic understanding of visual scenes have recently undergone rapid growth, making accurate object detection feasible in images and videos in daytime scenes. It is natural to raise the question of how to extend those sophisticated methods to other weather conditions and illumination conditions, and examine and improve the performance therein.

The semantic tasks are usually solved by learning from many annotations of real images. This strategy has attained great achievement for good weather conditions at daytime. However, collecting and labeling images for all other weather and illumination conditions are difficult. To overcome this problem, we skip this traditional paradigm and propose another route. Instead, we choose to progressively adapt the semantic models trained for daytime scenes to nighttime scenes, by using images taken at the twilight time as intermediate stages. The method is based on a progressively self-learning scheme, and its pipeline is shown in Fig 2. A method used to other weather conditions and illumination conditions and examine and improve the performance therein. This work would like to initiate the same research effort for nighttime.

## II. RELATED WORKS

### A. Semantic Segmentation

Semantic segmentation includes labeling each pixel of an image and therefore, keeping spatial information becomes extremely necessary. See Fig. 1 for an illustration. A neural network architecture used for scene parsing can be split into encoder and decoder networks, which are basically discriminative and generative networks respectively. State-of-the-art segmentation networks generally use categorization models which are mostly winners of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1] as their discriminator. The generator either uses the stored pooling indices from discriminator, or learns the parameters using convolution to implement up-sampling. Furthermore, encoder and decoder can be either symmetric (same number of layers in encoder and decoder with same number of pooling and unpooling layers), or they can be asymmetric. State-of-the-art semantic segmentation methods are primarily dependent on the recent advances of deep neural networks. As proposed by Long et al. [2], one can transform a classification Convolutional Neural Network (CNN) (e.g., AlexNet [3], VGG [4], or ResNet [5]) to a fully-convolutional network (FCN) for semantic segmentation. To train these advanced networks, a significant amount of dense pixel annotations must be collected in order to match the model capacity of deep CNNs. Simply, our goal is to take either a RGB color image (height×width×3) or a grayscale image (height×width×1) and output a segmentation map where each pixel contains a class label represented as an integer (height×width×1).



Fig. 1. An example of semantic segmentation

### B. Convolutional Neural Networks (ConvNets)

ConvNets were initially designed for image classification challenges, which consist in predicting single object categories from images. Long et al. [2] (FCN) firstly adapted known classification networks (e.g. VGG16 or GoogleNet) to perform end-to-end image semantic segmentation by making them fully convolutional and upsampling the output feature maps. However, directly adapting these networks results in coarse pixel outputs (and thus low pixel accuracy) due to the high down-sampling that is performed in the classification task to gather more context. To refine these outputs, Eduardo Romera, Luis M. Bergasa, Jose M. Álvarez and Roberto Arroyo [6] propose to fuse them with activation maps from shallower layers using skip connections. Kendall et al. [7] (SegNet) proposed to upsample the features with a large decoder segment that performs finer un-pooling by using the indices of the encoder's max-pooling blocks. Other works like [8] (DeepLab) have suggested to refine the coarse output by using CRFs, and works like [9] (CRFasRNN) recommended to combine them inside the convolutional architecture. However, relying on algorithms like CRF to refine segmentation highly increases the network's computational overload.

### C. Road Scene Understanding

Road scene understanding is an essential stage for applications such as supported or autonomous driving. Typical examples contain the detection of roads [10], traffic lights [11], cars and pedestrians [12], [13], and tracking of such objects [14], [15]. The purpose of this work is to adapt and extend the recent developed models for road scene understanding at daytime to nighttime, without manually annotating nighttime images.

## III. APPROACH

In our proposal, a semantic segmentation model on daytime images will be trained using the supervised learning model, and apply this model to a large dataset recorded at civil twilight time to create the class responses. The two subgroups of twilight are used: civil twilight and nautical twilight. Because the domain gap between daytime condition and civil twilight condition is relatively small, these class responses, along with the images, can then be exploited to fine-tune the semantic segmentation model so that civil twilight time can adapt this model. The same procedure is kept on through nautical twilight. And finally, nighttime images adapt the fine-tuned model. The stream of work on model distillation [16], [17], [18] is used for this learning method. Those methods either transfer management from complex models to simpler models for efficiency [16], [17], or transfer supervision from the domain of images to other domains such as depth maps [18]. We here transfer the semantic knowledge of annotations of daytime scenes to nighttime scenes via the unlabeled

images recorded at twilight time. It is called Gradual Model Adaptation method [19]. See Fig. 2 for an illustration.

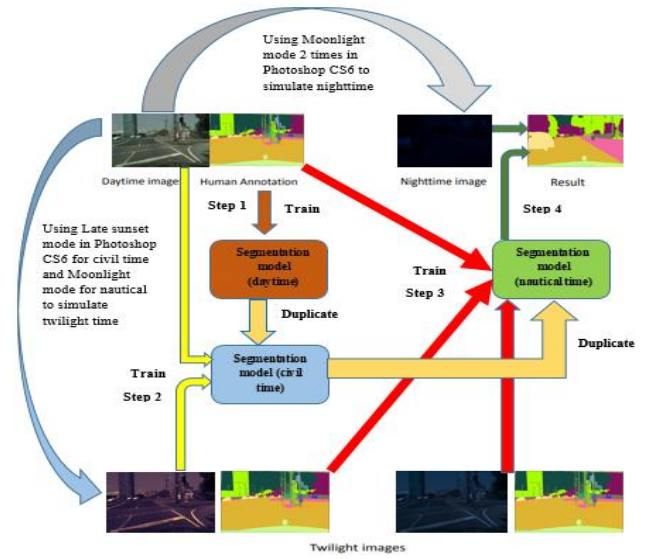


Fig. 2. The pipeline of our approach for semantic segmentation of nighttime scenes, by transfer knowledge from daytime scenes via bridge of twilight time scenes

Let us express an image by  $x$ , and indicate the image taken at daytime, civil twilight time, nautical twilight time and nighttime by  $x^0, x^1, x^2$ , and  $x^3$ , respectively. The corresponding annotation for  $x^0$  is provided and denoted by  $y^0$ , where  $y^0(m, n) \in \{1, \dots, C\}$  is the label of pixel  $(m, n)$ , and  $C$  is the total number of classes. Then, the training data contain labeled data at daytime  $D^0 = \{(x_i^0, y_i^0)\}_{i=1}^{l^0}$  and we use Photoshop CS6 software to convert daytime images to civil and nautical time, it means that we use same ground truth for daytime images for obtained civil and nautical images so that we get two labeled datasets for the two twilight categories:  $D^1 = \{(x_j^1, y_j^1)\}_{j=1}^{l^1}$  and  $D^2 = \{(x_k^2, y_k^2)\}_{k=1}^{l^2}$  where  $l^0, l^1, l^2$  are the total number of images in the corresponding datasets.

The Gradual model adaptation method consists of 4 steps. It is summarized below:

**Step 1:** train a segmentation model  $\Phi^0$  with daytime images and the human annotations:

$$\min_{\Phi^0} \frac{1}{l^0} \sum_{i=1}^{l^0} L(\Phi^0(x_i^0), y_i^0) \quad (1)$$

where  $L(\dots)$  is the cross entropy loss function

We use Pyramid Scene Parsing Network (PSPNet) [20] model to train a dataset with  $l^0 = 2300$  daytime images and its corresponding annotation and then optimizing the loss between the segmentation (predicted values -  $\Phi^0(x_i^0)$ ) which is created by PSPNet model and its annotation (groundtruth -  $y_i^0$ ). In this case, we use categorical cross entropy loss.

Categorical cross entropy will compare the distribution of the predictions with the true distribution, where the probability of the true class is set to 1 and 0 for the other classes. To put it in a different way, the true class is represented as a one-hot encoded vector, and the closer the model's outputs are to that vector, the lower the loss.

In detail, the Categorical cross entropy have the formula:

$$L(y, \hat{y}) = - \sum_{j=0}^M 1 \sum_{i=0}^N (y_{ij} * \log(\hat{y}_{ij}))$$

( $y$  corresponding to predicted value  $\Phi^0(x_i^0)$ ) and  $\hat{y}$  corresponding to ground truth -  $y_i^0$ )

where  $\hat{y}$  is the predicted value (It will in interval  $[0,1]$ )

$y$  is probability of class (It is set to 1 for true class and 0 for other classes)

$M$  is the number of classes

$N$  is the number of images

And about the metric, we use accuracy. This metric creates two local variables, total and count that are used to compute the frequency with which  $y\_pred$  matches  $y\_true$ .

**Step 2:** instantiate a new model  $\Phi^1$  by duplicating  $\Phi^0$ , and then retrain the semantic model on  $D^0$  and  $D^1$ :  $\Phi^0 \rightarrow \Phi^1$

And

$$\min_{\Phi^1} \left( \frac{\lambda^0}{l^0} \sum_{i=1}^{l^0} L(\Phi^1(x_i^0), y_i^0) + \frac{\lambda^1}{l^1} \sum_{j=1}^{l^1} L(\Phi^1(x_j^1), y_j^0) \right) \quad (2)$$

where  $\lambda^1$  is a hyper-parameter balancing the weights of the two datasets.

In this step, we use two datasets with the same scene but different time (one dataset at day time and another one at civil time) to train. It includes 2300 images for  $l^0$  and 2000 images for  $l^1$ .

**Step 3:** instantiate a new model  $\Phi^2$  by duplicating  $\Phi^1$ , and finetune (train) semantic model on  $D^0$ ,  $D^1$  and  $D^2$ :  $\Phi^1 \rightarrow \Phi^2$ , and then

$$\min_{\Phi^2} \left( \frac{\lambda^0}{l^0} \sum_{i=1}^{l^0} L(\Phi^2(x_i^0), y_i^0) + \frac{\lambda^1}{l^1} \sum_{j=1}^{l^1} L(\Phi^2(x_j^1), y_j^0) + \frac{\lambda^2}{l^2} \sum_{k=1}^{l^2} L(\Phi^2(x_k^2), y_k^0) \right) \quad (3)$$

where  $\lambda^1$  and  $\lambda^2$  are hyper-parameters regulating the weights of the datasets.

In this step, we use three datasets with the same scene but different time (first dataset at daytime, second one at civil time and the last one at nautical time) to train. It includes 2300 images for  $l^0$  and 2000 images for  $l^1$  and 2000 images for  $l^2$

**Step 4:** apply model  $\Phi^2$  to nighttime images to perform the segmentation:

$$\hat{y}^3 = \Phi^2(x^3).$$

Finally, we will predict for nighttime images after 3 above steps.

Where  $\lambda^1$  and  $\lambda^2$  are hyper-parameters regulating the weights of the datasets.

#### IV. DATABASE AND TRAINING PROCESS

##### A. Database

We use the Cityscape dataset [21] at daytime and using Photoshop CS6 software to convert from day to civil, nautical and night time images. In detail, we choose Late sunset mode

for civil time, Moon light mode for nautical and late moon mode 2 times for nighttime. It is illustrated in Fig.3 and Fig.4

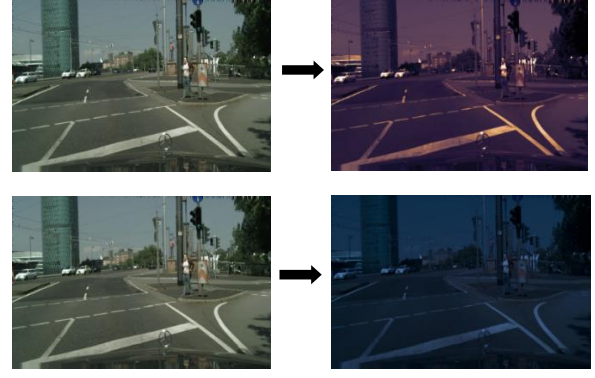


Fig. 3. Converting daytime images to twilight time and nautical time images using Photoshop CS6 Software

For testing images, we also use 200 modified images by Photoshop CS6 software.



Fig. 4. Converting daytime images to nighttime images using Photoshop CS6 Software

##### B. Training Process

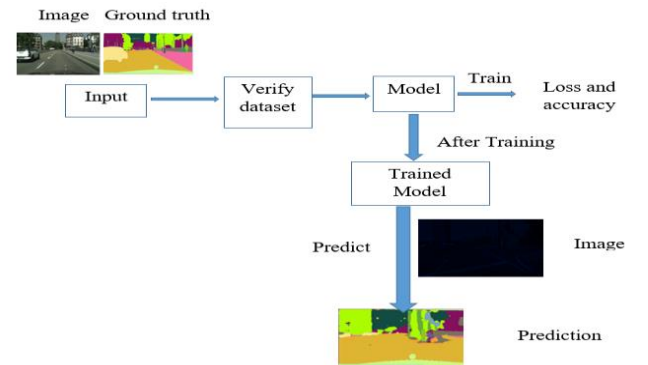


Fig. 5. Block diagram of algorithm

#### V. RESULTS AND EVALUATION

Our choice for experiments on semantic segmentation is the PSPNet [20]. In all experiments of this section, we use an Adadelta Optimizer. Our segmentation experiment indicates that using twilight time as a bridge is extremely effective,.

We evaluate mIoU of our method and compare it to the original segmentation model trained on daytime images directly. Using the pipeline described in Fig. 2, two models can be achieved, in particular  $\Phi^1$  and  $\Phi^2$ .

TABLE I. THE RESULTS FOR LAST EPOCH IN TRAINING AND EVALUATION METRIC OF GRADUAL MODEL ADAPTATION METHOD

Step	Loss	Accuracy	mIoU(%)
Step 1	0.2031	0.9401	46.82%
Step 2	0.1754	0.9459	46.96%
Step 3	0.1599	0.9486	46.71%
Step 4			44.9%



TABLE II. QUANTITATIVE RESULTS OF OUR METHOD AND ORIGINAL SEGMENTATION MODEL TRAINED ON DAYTIME IMAGES DIRECTLY

Method	mIoU(%)
Original method (without twilight)	36.71%
Our method (Gradual adaptation method)	44.9%

**Quantitative results:** The overall intersection over union (IoU) over all classes of the semantic segmentation are reported in Tables 1 and Table 2. The Table 1 presents results of all steps proposed in our method. We can observe that mIoU steadily decreases when brightness is reduced. The Table 2 shows that all variants of our adaptation method improve the performance of the original semantic model trained with daytime images. This is mainly due to the fact that twilight time falls into the middle ground of daytime and nighttime, so the domain gaps from twilight to the other two domains are smaller than the direct domain gap of the two. Also, it can be seen from the table that our method benefits from the progressive adaptation in two steps, i.e. from daytime to civil twilight, from civil twilight to nautical twilight. This means that the gradual adaptation reduce the domain gap progressively.

**Qualitative results:** We also expose multiple segmentation examples by our method and the original method in Figure 6. From this figure, one can show that our

method generally yields better results than the original method. For instance, in the first image of Figure 6, the original method misclassified some sky area as road. While improvement has been observed, the performance for nighttime scenes is still a lot worse than that for daytime scenes.

**Remarks:** After training 20 epochs, we can see that if the model are trained with 3 dataset including day, civil and nautical time, the prediction for nighttime images will be better than just using day or day and civil dataset. As the model is adapted one more step forward, the gap to the target domain is further narrowed. Data recorded through twilight time constructs a trajectory between the source domain (daytime) and the target domain (nighttime) and makes daytime-to-nighttime knowledge transfer feasible.

## VI. CONCLUSION

Semantic image segmentation is a key application in image processing and computer vision domain. This work has examined the problem of semantic image segmentation of nighttime scenes from a novel perspective. This method has proposed a novel method to progressive adapts the semantic models trained on daytime scenes to nighttime scenes via the twilight time

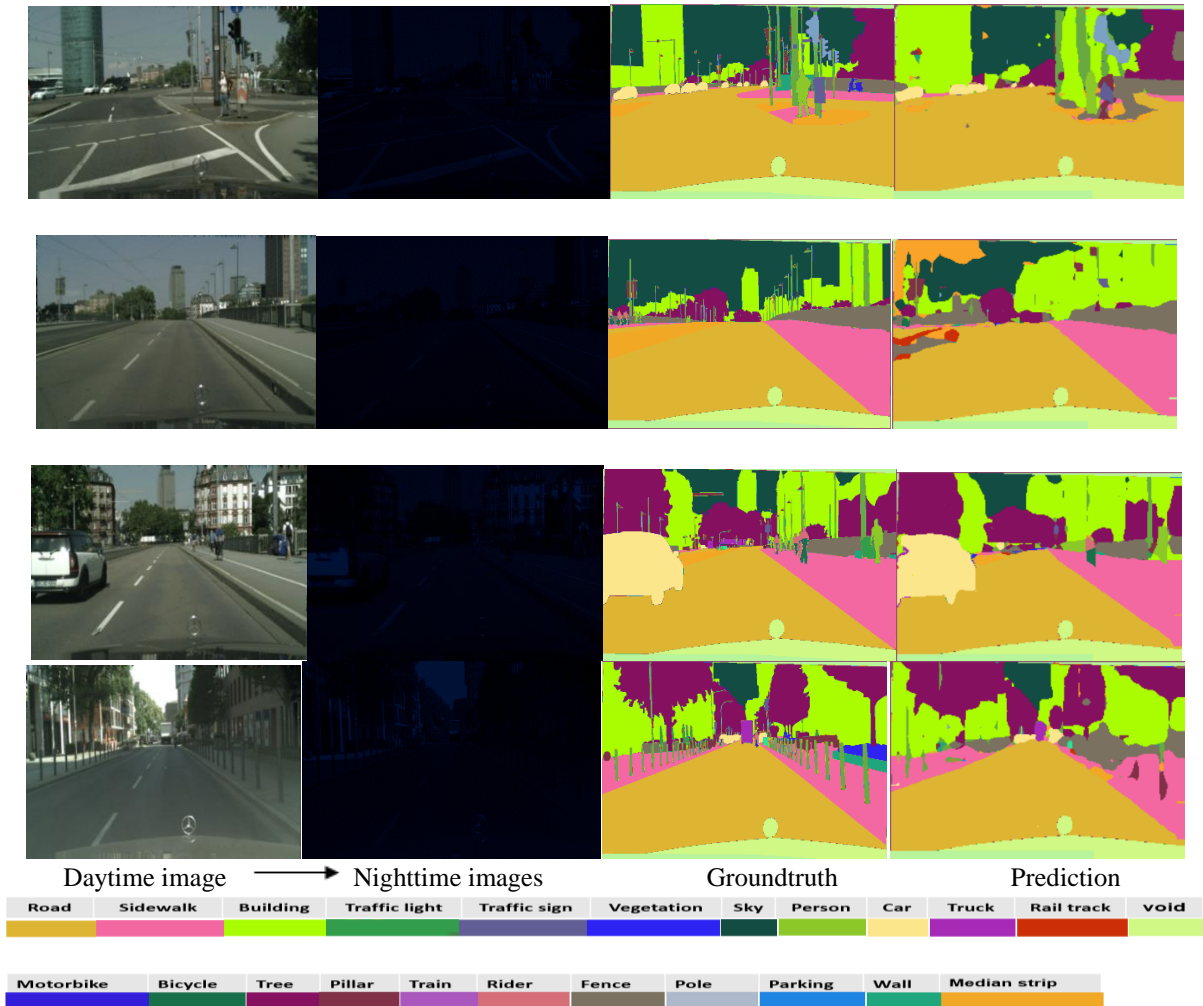


Fig. 6. Qualitative results for semantic segmentation on nighttime time scenes



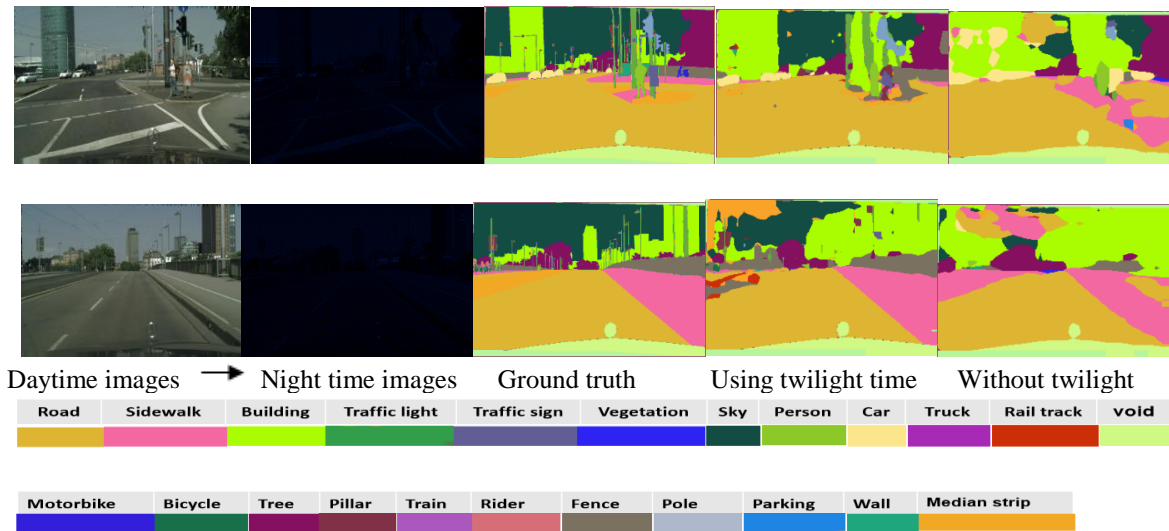


Fig. 7. Comparison qualitative results for semantic segmentation on nighttime time scenes when using twilight time as bridge and without twilight time

This work has researched the problem of semantic image segmentation of nighttime scenes. This paper has proposed a novel method to progressively adapt the semantic models trained on daytime scenes to nighttime scenes via the bridge of twilight time — the time between dawn and sunrise, or between sunset and dusk. Data recorded during twilight times are further grouped into two subgroups for a two-step progressive model adaptation, which is able to transfer knowledge from daytime to nighttime in an unsupervised manner. The experiments show that our method is effective for knowledge transfer from daytime scenes to nighttime scenes without using human supervision.

While improvement has been observed, the performance for nighttime scenes is still a lot worse than that for daytime scenes. Nighttime scenes are indeed more challenging than daytime scenes for semantic understanding tasks. There are more underlying causal factors of variation which requires either more training data or more intelligent learning approaches to extricate the increased number of factors. Introducing a sufficient amount of human annotations of nighttime scenes will for sure improve the results.

#### ACKNOWLEDGMENT

This work was carried out under capstone project collaboration between by the Faculty of Advanced Science and Technology, The University of Danang - University of Science and Technology and Department of Electrical Engineering- National Chung Cheng University. Specially thanks to the SmU2SmC Research Team at UD-DUT which provided insight and expertise that greatly assisted the research in this paper.

#### REFERENCES

- [1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recog. (CVPR)*, 2015, pp. 3431–3440.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [6] E. Romera, J. M. Álvarez, L. M. Bergasa and R. Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, Jan. 2018.
- [7] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixelwise labelling," *arXiv preprint arXiv:1505.07293*, 2015.
- [8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.
- [9] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conf. on Computer Vision (ICCV)*, 2015, pp. 1529–1537.
- [10] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz. Recent progress in road and lane detection: A survey. *Mach. Vision Appl.*, 25(3):727–745, Apr. 2014.
- [11] M. B. Jensen, M. P. Philipsen, A. Mgelmoose, T. B. Moeslund, and M. M. Trivedi. Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):1800–1815, July 2016.
- [12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [13] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018.
- [14] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795, 2013.
- [15] S. Manen, M. Gygli, D. Dai, and L. Van Gool. Pathtrack: Path supervision for efficient video annotation. In *International Conference on Computer Vision (ICCV)*, 2017.
- [16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [17] D. Dai, T. Kroeger, R. Timofte, and L. Van Gool. Metric imitation by manifold transfer for efficient vision applications. In *CVPR*, 2015.
- [18] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

- [19] D. Dai and L. V. Gool, "Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018, pp. 3819-3824
- [20] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6230-6239.
- [21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

# Proposed Smart University Model as a Sustainable Living Lab for University Digital Transformation

Tuan V. Pham

University of Science and Technology  
The University of Danang  
Danang, Vietnam  
pvtuan@dut.udn.vn

Anh Thu T. Nguyen

University of Science and Technology  
The University of Danang  
Danang, Vietnam  
ntathu@dut.udn.vn

Thanh Dinh Ngo

University of Science and Technology  
The University of Danang  
Danang, Vietnam  
ndthanh@dut.udn.vn

Duy H. Le

University of Education and  
Technology The University of Danang  
Danang, Vietnam  
lhduy@ute.udn.vn

Khai C.V. Le

TAPIT Engineering Co., Ltd.,  
Danang, Vietnam  
vinhkhai95@gmail.com

Thuong H.N. Nguyen

TAPIT Engineering Co., Ltd.,  
Danang, Vietnam  
nhathuongqn@gmail.com

Huy Q. Le

University of Science and Technology  
The University of Danang  
Danang, Vietnam  
lqhuy@dut.udn.vn

**Abstract**—Recently, there is an emerging and rapidly growing research on smart universities due to the impact of advanced technologies such as artificial intelligence, IoT technologies and big data to enhance performance of teaching activities and university management systems. The motivation of this paper is to study university digital transformation and analyze smart university framework. The study begins with a description of trends in initiating various concepts, challenges as well as opportunities on development of smart universities. The second contribution focuses on expressing a systematic approach of designing the smart university concept, that separates its application domains and covers its technological system. A smart university model is then proposed which includes five domains: IoT ecosystem, smart infrastructure, smart applications and services, smart teaching and learning, data-driven smart analysis. Lastly a proposal of establishing a smart university as a living laboratory will be discussed. Several pilot projects which have been implemented in the UD-DUT testbed lab will be further analyzed, such as employee attendance management system, smart office energy management, campus public LoRA network. In addition, possible challenges and future works regarding this novel smart university framework will be explored.

**Keywords**—Smart University, Smart Campus, IoT Ecosystem, Smart Teaching and Learning, Digital Transformation

## I. INTRODUCTION

### A. Trends in Smart University

Recently, research and development of smart universities (SmU) and related topics have become the main themes of various pioneering national and international studies, events and projects, governmental and corporate initiatives, institutional agendas and strategic plans. A study in [1] presents outcomes of an ongoing project aiming at a systematic literature review and creative analysis of professional publications in those areas. The outcomes enable researchers to identify the well-thought ideas, approaches and best practices for the next evolutionary generation of the SmU.

The Smart University Foundation in [2] which is a global membership-based initiative on “Smart University” extensively focuses on eight domains: Smart Campus, Smart Education, Smart Research, Smart People, Smart Quality, Smart Recruitment, Smart Governance and Smart Influence.

A study in [3] shows that the use of intelligent technologies as a link between people and their university environment will soon change the way the individual interacts with the university environment. The approach described in [4] defined SmU as a platform that acquires and delivers foundational data to drive the analysis and improvement of the teaching and learning environment. The study in [5] presents “Intelligent University” is a concept that includes a comprehensive modernization of all educational processes. They argue that intelligent education is able to reach by the use of ICT tools for most activities of lecturers leading to a completely new variety of processes and results of educational, research and other activities. L.F. Kwok defines in [6] that SmU is a new paradigm of thinking pertaining to a holistic intelligent campus environment which encompasses at least, but not limited to, several themes of campus intelligence. The authors in [7] argue that universities should become smarter places where knowledge is shared between employees, teachers, students, and all stakeholders in a seamless way. Besides, designers of SmU should pay more attention to the maturity of smart features of SmU that may occur on various levels of smartness, such as adaptation, sensing, inferring, self-learning, anticipation, and self-organization [1].

### B. Challenges and opportunities

The recent disruptive technologies open doors for researchers to investigate and promote the campus services using innovative solutions. In addition to recent works focusing on smart learning environment, intelligent buildings, efficient energy saving, there has been much attraction on developing smart campus applications and services which includes transportation, parking and traffic, waste and

hygiene, water and air quality, safety and security, social and sport activities. However, the SmU market is facing technical, financial, and political obstacles which are similar to those faced by the smart city market [8].

The technical challenges can be observed from the following perspectives: safety, security, privacy, interoperability, standardization, and configuration [9]. It is necessary to consider impact of new technologies to people on campus and the public before adopting them to university applications and services. Requirement on security and privacy should be carefully considered due to the heterogeneous education environment of education, research and public services. The interoperability and standardization in such heterogeneous environments would maximize the users benefits. These allow the evaluation and comparison of devices, promote the devices introduction to the ecosystem, and allow competition between manufacturers. Lastly, adaptation of IoT technology on campus would result in using thousands of sensors, actuators, and other “things”, which can be an enormous burden to configure them manually.

On the financial side, the shrinking of investments on public services and limited resources provided to universities are preventing the smart university from becoming reality, regardless of immature experiences [10]. The political challenges should also be considered before moving to a SmU. Collaboration between different colleges and departments, re-engineering of business processes, and opposition of anti-tech employees are potential obstacles which need to be resolved. Generating a model of intelligent maturity can be seen as an evolutionary approach for a traditional university to progress to several levels of maturity of a SmU [11].

Therefore, there is an essential need to develop a new conceptual SmU model together with identification of associated systems, features, technologies and smartness. The rest of the paper is organized as follows. Section 2 expresses related works in different domains of smart university. Section 3 proposes a systematic approach of smart university architecture. Section 4 discusses the living laboratory as an approach for transition to smart university. Finally, section 5 concludes the paper and highlights some future research directions.

## II. RELATED WORKS

### A. IoT Ecosystem

A smart campus should ensure many IoT applications rely on a wide range of different communication standards and should be able to talk together at the same time. The paper in [12] presents WiFi devices that use MQTT protocol to communicate through the network. In order to auto-configure and manage MQTT devices, IoT gateway is introduced in [13]. The authors in [14] promote smart campuses and smart cities based on wireless communication technologies including LoRa and NB-IoT. As shown in [15], LPWANs communication technology for IoT applications is a potential solution to address the mobility of the things and the connectivity.

It is obvious that the more establishment of IoT technologies the more choices of high quality and functionality products customers can have. However, in order to use new communication standards, users have to replace the whole previous system, from central processors to sensors and actuators. Besides, for the same communication protocol but

developed by different companies, it is not easy to integrate them together. Thus, combining the advantages of sensors from various companies together is impossible since different communication protocols cannot understand each other. Conversely, diversity of IoT products results in inconvenience for customers when they have to pay more money to exploit advantages of latest technology from different companies. In other words, it is hard to integrate modern technologies into a current system. For that reason, it is necessary to have a so-called IoT ecosystem gateway that can solve all of those problems.

### B. Smart Infrastructure

A smart classroom relates to the optimization of content presentation, convenient access of learning resources, deep interactivity of teaching and learning, contextual awareness and detection, classroom layout and management etc. The study in [43] proposed a SMART model that covers Showing, Manageable, Accessible, Real-time Interactive and Testing features. From that, a set up the “iSMART” system for practical operation has been further developed in [16], which is composed of six different modules: infrastructure, network sensor, central management, augmented reality, real-time recording, and ubiquitous technology.

Deploying AI-based robot products at schools has been carried out and achieved remarkable results in many countries. ZenoBot is an AI teaching assistant product being tested successfully with students at Trinity Lutheran College [17]. The intelligent robot Jill Watson, one of the nine teaching assistants of Ashok Goel, who was able to answer up to 97% questions of the Knowledge-Based Artificial Intelligence online [18].

Smart offices make use of technology to automate everyday work to truly optimize processes. According to a study by [19], a model of smart office has been proposed with necessary elements in a Building Management system, including: water, lighting, heating, power, health, air conditioning, room, security, access. In [20], the authors present a comprehensive smart office system concentrating on door-access, lighting, illuminating, ventilating, heating, and reconfiguration is designed in order to save energy and promote the satisfactions of the employees.

The research in [21] develops the outline of a new concept of the smart library, which can be described in four dimensions: smart services, smart people, smart place, and smart governance. The Technical Information Center of Denmark has been built according to the living lab model to become a space for students and graduate students to develop, test and carry out research projects. Their smart library model is defined with 4 elements: personal comfort, technological playground, open data repository, environment and economic sustainability [22]. The study in [23] presents a “Smart Laboratory” system of three-layer structure models consisting of Perceptual, Network and Application to meet requirements of university research laboratories.

### C. Smart Application and Services

In the concepts of SmU, recent studies rely on smartphone or computer applications to monitor human activities based on location, mobility, and social interaction [24]. Other AI-based applications have been developed to enhance university management performance, such as facial recognition, fingerprint identification, people counting, working time

estimation in departments, activity determining of crowded places. There applications are carried out to manage human resources, optimize student activities, increase quality of services in university. Some studies focus on safety and security by combining different sensors with video recording devices [25], locating and detecting illegal access or abnormal behavior to guarantee security in university [26]. Moreover the valuable equipment are mounted to the tracking devices to avoid being lost and stolen. Protection of private data of users using wireless networks in university is also considered in [27]. The issue of green environment on university campus was also studied and deployed as irrigation systems, smart parking [28].

Many resources in university campuses can be monitored through different IoT and AI systems such as garbage collection monitoring [29], water management [30] energy consumption management [31] and other solutions for sustainable development. Especially, some recent research is realized with a target of analyzing health in the campus [32], monitoring stress and concentration level of students on lessons [33]. Some studies about data logger, which is a device to collect, monitor and transfer data automatically has been applied to monitor environment parameters such as water level, water quality, air quality. It can be used to automatically control irrigation - pump system, electric system in university campus.

#### D. Smart Teaching and Learning

There have been many studies recently focusing on how to support students to learn in real-world contexts in smart ways. Smart educational technology is becoming one of popular approaches to reach the goal. Though there are various ways of understanding and implementing e-learning, a definition in [34] simply emphasizes the use of computer and Internet technologies to deliver a broad array of solutions to enable learning and improve performance. A comparison of different factors of ICT and e-learning in several countries has been carried out in [35]. The research results showed that due to the extensive use of ICT in e-learning, more and more legislative activities are being undertaken to ensure a flexible legal framework for the implementation of these technologies in education at different levels.

The research in [36] presented a concept of intelligent learning environments that can be considered as a learning environment supported by technology. They introduce adaptation and provide appropriate support at the right places and at the right time for individual student needs, which can be determined by analyzing their educational behavior as well as the online and real context in which they are located. In [37], the role of e-portfolio for smart life-long learning is described. The paper implies a smarter curriculum that facilitates self-regulated learning of learners through the introduction of personal development e-Portfolio; assists learners in planning their development path and reflecting upon their own learning. Recently, a research in [38] shows that in the near future, smart pedagogy will be actively deployed by leading academic institutions in the world for teaching of local and remote students. Several research findings suggest that preservice teachers may need more guidance, modelling and collaboration to develop a better understanding of technology-based pedagogy from their own practice so that they can synthesize their constructivist orientation and student-centered teaching approaches.

#### E. Data-driven Smart Analysis Management

The combination of big data with analytical and statistical applications in management platforms has become a trend in modern higher education [39]. The study in [40] confirms that big data certainly plays an important role for management and analytical processes in higher education. Administrators and teachers can check student performance and feedback using data gained from course evaluation, class discussion, or other learning activities. These data which can be collected in real time or near real time in the course time or after course time, help teachers to make decisions on quality improvement. As addressed in [39], these kinds of data analysis processes can be considered as fundamental tools to provide deeper understanding and useful information for making wise decisions.

In the context of higher education, Big Data has been applied in analyzing a wide range of administrative data and data collected from operations of university itself such as academic programming, research, teaching and learning [41]. Big data in higher education management also includes physical database systems that can store large amounts of data such as student information, or university operation activities, or specific activities about learning and teaching. Others researchers showed that in order to improve productivity at higher education, universities must include analytical tools in the system [42].

### III. PROPOSED SMART UNIVERSITY MODEL

#### A. Proposed smart university model

**Smart university model.** There is a need to develop a SmU model that reserves a common framework, can adopt diverse technologies to meet demands of stakeholders and is flexible enough to accommodate diverse cases as identified above. In this study, a Smart University model is proposed with the following taxonomies: *key features*, *systems*, *components* and *smartness levels*. The proposed SmU architecture is shown in Figure 1.

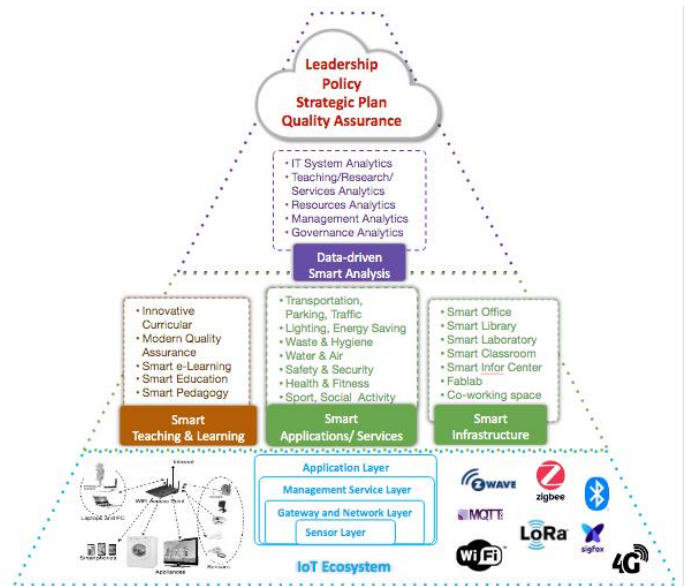


Fig 1. Proposed smart university architecture.

**Key Features.** Improved student engagement and experiences; better proximity to skills of the digital economy;



better operational efficiencies; better learning outcomes; contemporary industry partnerships; connected community.

**Systems.** Layered structure design and implementation; interoperability; scalability; flexibility; fault tolerant; availability, manageability & resilience; standard-based criteria; technology and/or vendor independence.

**Components.** IoT ecosystem; smart infrastructure; smart applications and services; smart teaching and learning; data-driven smart analytics.

**Smartness levels.** The Smart Maturity model developed in [1] can be viewed as an evolutionary approach for a traditional university to progress to various levels of the maturity of smart university.

### B. Smart university components

**IoT Ecosystem Network.** As discussed above, there are diversities of smart things including learning management system, library, building management system, security system, etc. in smart campuses. This means that there are a lot of “things” so called heterogeneous devices needed to work together. As nature, they use different protocols from different technologies provided by various vendors. Therefore, an IoT ecosystem gateway should be developed to create all in one smart campus package.

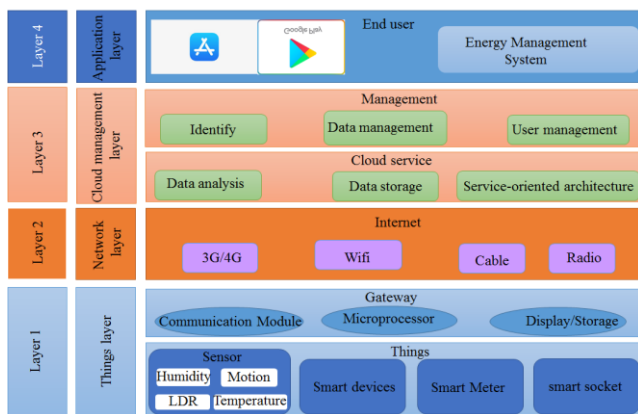


Fig 2. Proposed IoT eco-gateway Architecture.

In our study, an IoT eco-gateway has been proposed with 4 layers as depicted in Fig.2. This gateway supports a variety of communication protocols such as Wi-Fi, Bluetooth, Z-wave, LoRa, etc.. Things layer consists of 2 sublayers of all things such as sensors, actuators, smart sockets, electronic meters responsible for measuring, collecting data and controlling devices. household appliances. In addition, the gateway layer consists of a microprocessor, communication modules, storage, and display that allow the universal layer to be connected. Network layer allows connecting the device layer to the cloud management layer. Cloud management layer builds cloud storage services to manage user data and system historical data. Application layer provides services to the customers in the smart campus including control, monitoring, and acquisition.

**Smart Infrastructure and Application.** Our aforementioned research and application results in Smart University have contributed to the following smart rules in building intelligent applications and services in the university:

**Smart applications.** Any applications which are developed for SmU should meet the following demands of scalability,

flexible responsiveness, low-power distance communications, new technologies integration.

**SMART things.** The SMART models in [16, 43] should be applied and tracked in design - operation - maintenance - improvement processes for developing smart universities. Basically, all six different modules of a SMART thing should be covered: infrastructure, network sensor, central management, augmented reality, real-time recording, and ubiquitous technology.

**Smart Teaching and Learning Framework.** It is important to propose a framework of smart teaching and learning to serve a broader educational development agenda than only technical interoperability. It should be noticed that smart education with technologies will confront many challenges, such as pedagogical theory, educational technology leadership, teachers learning leadership, educational structures and educational ideology.

**Smart Educational Technology.** It should be focused on innovative technology-based constructivist learning approaches to engage students in student-centered learning such as: flipped classrooms, learning-by-doing including usage of virtual laboratories, collaborative learning, adaptive teaching, personal enquiry based learning and other innovative strategies.

**Smart Pedagogy.** From our experiences, the smart pedagogy framework studied in [44] shows their potential in supporting students reaching future soft skills. There are four groups of smart pedagogy: class-based differentiated instruction, group-based collaborative learning, individual-based personalized learning and mass-based generative learning.

**Smart Quality Assurance.** The studied outcomes from different research on smart education prove that the learning and teaching strategies in smart pedagogy support identified “smartness” levels and smart features such as adaptivity, sensing, inferring, anticipation, self-learning, and self-organization. Therefore, it is needed to develop a systematic quality assurance of its functionalities, processes, outputs, outcomes and impact.

**Data-driven smart analytics:** The architecture consists of 2 main components: Analytical tools and Centralized database.

**Analytical tools.** The analysis tools consist of four main components which are hierarchized into different levels, and accessed by users with proper permission depending on their roles and duties in the university. Institutional analytics is the analytical tool with the highest hierarchy. It provides macro-level analytical tools related to the school's policies, operating models, and provides high-level statistics to help make operational decisions for policies. Information Technology System Analytics integrates data at a lower level and lower hierarchy from many different systems being operated in the school such as student information systems, academic management systems, alumni update and storage systems, experimental equipment management systems, etc. Academic/ Program analytics encapsulate all activities in higher education institutions that focus on analyzes on the learning process, relationship between learners, content, organization and instructor, or the relationships in academic research effectiveness. The academic data will focus on employees involved in the academic or academic program, or

faculty or administrators. Learning analytics provides new tools that are more effective for analyzing and manipulating teaching data. Data may be student behaviors, such as number of visits to a specific lesson, or individual student feedback about the subject or about the different parts of the subject.

**Centralized database.** Centralize data can be classified into 6 main data types: Data related to school administration and policies; Data related to the school's research activities; Data related to students as personal information, scores, etc.; Data related to teaching and learning such as student feedback, subject performance surveys, attendance, etc.; Data related to school activities such as facilities, resource management, etc.; Data related to training programs, etc

#### IV. TRANSFORMATION INTO SMART LIVING LAB

##### A. Smart Campus as a Living Lab

**Living Laboratory.** One of the effective approaches to fulfil the above requirements is to develop campus as a living laboratory. In order to transform traditional universities to smart universities, we propose a set of simultaneous requirements:

- i) Taking into account personal expectations and needs of related stakeholders;
- ii) Preparing specialists who have professional skills and habits in a smart society;
- iii) Implementing cutting-edge technologies, innovative sustainability transformations, best academic practices and current trends in management and leadership for a better future of higher education.

The research report in [45] is about how universities can help solve the challenges faced by smart cities and towns. University and city staff should build effective community engagement into demanded smart projects. As being concluded in [46], smart cities require smart universities that follow sustainable principles in all their activities and enable others to do the same. A study in [10] considers a university campus as a small city which serves as a test bed for the integration of techniques that make up a smart campus. This concept allows identification of the characteristic features of a campus based on those of a city: complexity, diversity, uncertainty, sustainability.

**The Living Smart Campus at The University of Twente.** The Living Smart Campus program provides an environment for working on complex social issues that call for scientific solutions [47]. The campus living environment offers a unique setting in which developed products and systems are prepared, tested and improved before they are introduced into society. This living lab creates external visibility and collaboration with business; support faster implementation process and integrated approach.

**Smart Campus Project at University of Glasgow.** The Smart Campus project is an exciting opportunity which create a dynamic experience for staff and students across the University's estate [48]. Its purpose is to develop the University as an international focus for future city best practice. The program integrates enhanced research projects within the new campus: Urban infrastructure innovation; City, district systems; Sensors and sensor systems; Enhanced building management; Transport policy; Enhanced learning

experience; Staff and student health and well-being; Big data analytics.

**Central University of Tamil Nadu (CUTN).** There has already been a variety of work undertaken at CUTN to support the development of a Smart Campus: a) The IoT integrates sensors, controllers, machines, people, and things in a new way to realize intelligent identification, location, tracking, monitoring, and management; b) Big data technologies including mass data acquisition, mining, storage and processing make its management services to a higher level; c) Knowledge management which is the planning and management of knowledge, knowledge creation process and knowledge application.

**Smart University in Smart City.** It is important for universities to be measured on performance metrics related to economic, social and cultural sustainability. The development of "smart cities" depends strictly on smart universities, whose action influences employment, organizational and educational management, the development of human and social capital associated with economic benefits [3]. The majority of European countries are taking this path and are focusing on such aspects. The research report in [45] is about how universities can help solve the challenges faced by smart cities and towns. University and city staff should build effective community engagement into demanded smart projects. Successful smart city programs balance quick, visible wins while taking a long-term perspective, recognizing that effective change can be a slow and sometimes messy process. As being concluded in [46], smart cities require smart universities that follow sustainable principles in all their activities and enable others to do the same. A study in [10] considers a university campus as a small city which serves as a test bed for the integration of techniques that make up a smart campus. This concept allows identification of the characteristic features of a campus based on those of a city: complexity, diversity, uncertainty, sustainability. The components in this work are related to three main axes: IoT, cloud computing and big data.

##### B. Pilot projects

**Automatic Monitoring System.** Control and monitoring systems for electrical devices in classrooms have been implemented using Bluetooth Mesh, LoRa and Wifi technology at the University of Danang – University of Science and Technology (DUT). All information about users, managers and classroom is stored and processed at the webserver as shown in Figure 3 below.

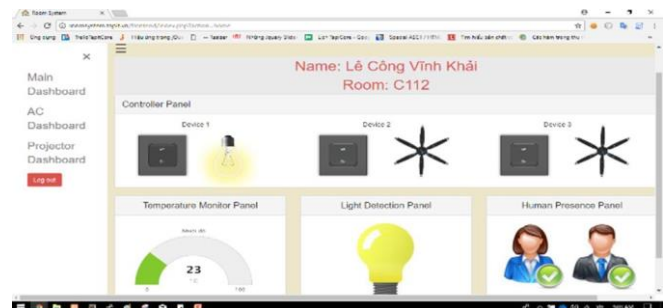


Fig 3. Web application of the automatic monitoring system

The LoRa gateway uses RFM95W module and W5100 Ethernet module. The Bluetooth Mesh gateway is designed using wireless MCU nRF52832 and ESP32 Wifi module. The

message from the gateway is sent to the server using the MQTT protocol. Nodes are designed to collect temperature, light, and motion sensors and control the devices via switch board and transmit infrared codes infrared signals. The system has been tested with 400 data transmissions, the results show a 100% success rate and an average transmission time of < 1 second.

**Attendance management system (AMS).** An automated employee AMS has been developed, installed and implemented at DUT building [49]. Figure 4 shows its logical scheme. The face recognition based AMS system has been trained and tested on a self-built database consisting of DUT employees. Various algorithms and deep learning models have been designed for building the face recognition module. The recognition accuracy reaches more than 98% on the well-matched testing set and more than 82% on the high-mismatched testing set, respectively. An web application has been integrated to the system to provide administrative information such as: attendance information (name, unit, position, image); statistics of staff working frequency; alarm announcement for detected strangers.

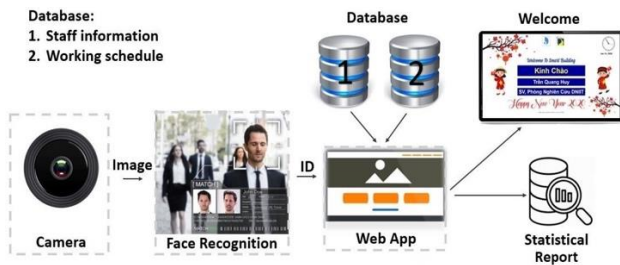


Fig 4. Automated employee attendance management system

**Datalogger Device.** A datalogger device has been developed at DUT to collect, control, display and transmit data automatically to server via Internet with HTTP and FTP protocols. The device depicted in Figure 5 uses a powerful ARM Cortex-M core microcontroller that has been designed with a wide range of features, flexible application deployment through isolated digital input/output, analog voltage input and 4-20mA current input standards. In addition there are communication standards such as Ethernet, 3G/4G, RS232, RS485 integrated in this module.



Fig 5. Datalogger device

A test on operation of 02 devices for 30 consecutive days with water quality monitoring application with a frequency of every 5 minutes had been carried out. The obtained results

showed a total packet of 103,680 packets per station with a 100% success rate. This shows that the device operation is very stable and effective in management, operation and maintenance.

**LoRa network for Smart DUT Campus.** Two industrial LoRa gateways were installed in Danang city, one on top of the administration building of DUT and another on top of Danang's Software Park located in the city center. This establishes a zone around DUT campus covered with LoRa network for anyone to develop their IoT applications for Smart Campus and Smart City. In this experiment [14], we used sensor node including B-L072Z-LRWAN1 board in combination with X-Nucleo-IKS01A1 shield to collect environmental data. The MEMS environmental sensor is located in a moving vehicle at a speed of 40 km/h with a probability of transmitting data to two gateways as shown in Figure 6. Our test results demonstrates that the whole DUT campus as well as the surrounding zone about 6 km of diameter around the campus are readily covered by the LoRa network. The farthest distance of data exchange is over 10 km. With these initial results, the LoRa network is a potential candidate for long-range IoT applications in smart campus.

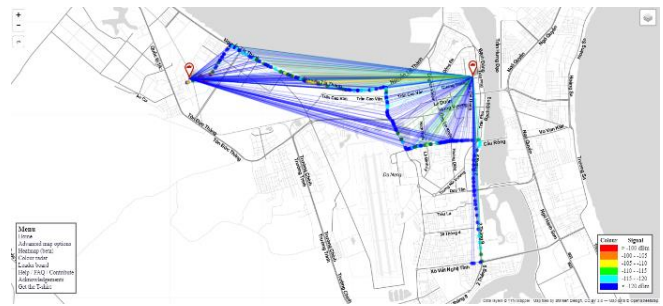


Fig 6. LoRa coverage range with mobile sensor node testing

## V. CONCLUSION

In this paper, we have outlined our deep research towards the common framework of smart university and its benefits as well as challenges. We went through literature of smart universities, practices of selected universities around the world. Vision and framework of a smart university has been articulated with five domains systematically integrated and interacted among in this smart university architecture. It encompasses IoT Ecosystem, Smart Infrastructure, Smart Applications and Services, Smart Teaching and Learning, and Data-driven Smart Analytics. Lastly, we proposed a living lab with key features and pilot projects which have being executed in our campus.

As a conclusion, a smart university must ultimate technological solutions to foster collaboration among individuals, create conditions for effective study and development, face urgent challenges, create innovation, and prepare new generation specialists, new directions for functioning in a new smart society and smart environment. It is important that university needs to determine its smart vision in perhaps five years to ten years, then planning back to determine phases and steps to achieve the vision. When all of the people, devices, and applications on campus share a common technology infrastructure, they can interact with each other to enable experiences and efficiencies that weren't possible before. Universities should better become smarter places where knowledge is shared between employees, teachers, students, and all stakeholders in a seamless way.



## ACKNOWLEDGMENT

This work was supported by The University of Danang, University of Science and Technology through the research projects with code number T2019-02-40, T2019-02-41. Specially thanks to the "Smart University towards Smart City - SmU2SmC" Research Team at UD-DUT which provided insight and expertise that greatly assisted the research in this paper.

## REFERENCES

- [1] Heinemann, Colleen, and Vladimir L. Uskov. "Smart university: Literature review and creative analysis." in *Smart universities: concepts, systems, and technologies*, Vol. 70, Springer, 2018.
- [2] SUF, About Smart University Foundation, 2019 <https://s-u-f.org/>
- [3] Nuzzaci, Antonella & La Vecchia, Loredana, "A Smart University for a Smart City". *International Journal of Digital Literacy and Digital Competence*. 3. 16-32. 10.4018/jdlc.2012100102, 2012.
- [4] C. S. Sauer, T. Sakur, S. Oussena and T. Roth-Berghofer, "Approaches to the use of sensor data to improve classroom experience," *eChallenges e-2014 Conference Proceedings*, Belfast, pp. 1-9, 2014.
- [5] Tikhomirov, V., Dneprovskaya, N., "Development of strategy for smart University", *Open Education Global International Conference*, Banff, Canada, 22–24 April 2015.
- [6] Kwok, L.F. "A vision for the development of i-campus", *Smart Learning Environments*, *Springer Open Journal*, 2:2, Springer, 2015.
- [7] Maresca, Paolo & Coccoli, Mauro & Guercio, Angela & Stanganelli, Lidia. "Smarter universities: A vision for the fast changing digital era". *Journal of Visual Languages & Computing*. 25. 1003-1011. 10.1016/j.jvlc.2014.09.007, 2014.
- [8] Mahmood, Zaigham, ed. "Smart Cities: Development and Governance Frameworks". Springer, 2018.
- [9] A. Alghamdi et al., "Survey toward a smart campus using the internet of things," in *Proc. IEEE FiCloud*, Vienna, Austria, 2016.
- [10] Villegas-Ch, W.; Palacios-Pacheco, X.; Luján-Mora, S. "Application of a Smart City Model to a Traditional University Campus with a Big Data Architecture: A Sustainable Smart Campus". *Sustainability* 2019, 11, 2857, 2019.
- [11] Rico-Bautista, Dewar. "Conceptual framework for smart university". *Journal of Physics: Conference Series*. 1409. 012009. 10.1088/1742-6596/1409/1/012009, 2019.
- [12] A. Bhatt, J. Patoliya, "Cost effective digitization of home appliances for home automation with low-power WiFi devices," in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, 2016.
- [13] S.-M. Kim, H.-S. Choi, W.S. Rhee, "IoT home gateway for auto-configuration and management of MQTT devices," in *IEEE Conference on Wireless Sensors (ICWiSe)*, Melaka, 2015.
- [14] Ngo Dinh Thanh, Fabien Ferrero, Doan Quang Vinh, Pham Van Tuan, "Industrial LoRaWAN network for Danang city: solution for long-range and low-power IoT applications", in *International Conference on Research in Intelligent Computing in Engineering (RICE 2020)*, Binh Duong, Vietnam, 2020.
- [15] W. Ayoub, A. E. Samhat, F. Nouvel, M. Mroue and J. Prévotet, "Internet of Mobile Things: Overview of LoRaWAN, DASH7, and NB-IoT in LPWANs Standards and Supported Mobility," *IEEE Communications Surveys & Tutorials*, pp. 1561-1581, 2019.
- [16] Song, Shuqiang & Zhong, Xiaoliu & Li, Haixia & Du, Jing & Nie, Fenghua, "Smart Classroom: From Conceptualization to Construction". 330-332. 10.1109/IE.2014.56, 2014.
- [17] "ZenoBot to support teacher only", *Australian Teacher Magazine*, pp. 42, March 2018.
- [18] Siau, Keng & Ma, Yizhi. "Artificial Intelligence Impacts on Higher Education", 2018.
- [19] CBS Interactive Inc., "How to optimize the smart office", 2018.
- [20] H. Li, "A novel design for a comprehensive smart automation system for the office environment," *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, Barcelona, pp. 1-4. doi: 10.1109/ETFA.2014.7005267, 2014.
- [21] Joachim Schöpfel, "Smart Libraries", September 2018.
- [22] DTU Library, Technical University of Denmark. "SMART library", June 2016.
- [23] Yan, Zi Qi, et al. "Research on the Structure of Smart Laboratory Based on the Internet of Things Technology." *Applied Mechanics and Materials*, vol. 427–429, Trans Tech Publications, Ltd., pp. 2605–2608, September 2013.
- [24] Miao, C.; Zhu, X.; Miao, J. "The analysis of student grades based on collected data of their Wi-Fi behaviors on campus". In *Proceedings of the IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, China, 14–17 October 2016.
- [25] Zhang, J. "Spatio-Temporal Association Query Algorithm for Massive Video Surveillance Data in Smart Campus". IEEE Access, 2018.
- [26] Liu, K.; Warade, N.; Pai, T.; Gupta, K. "Location-aware smart campus security application". In *Proceedings of IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, San Francisco, CA, USA, 4–8 August 2017.
- [27] Zhang, L.; Oksuz, O.; Nazaryan, L.; Yue, C.; Wang, B.; Kiayias, A.; Bamis, A. "Encrypting wireless network traces to protect user privacy: A case study for smart campus". In *Proceedings of the IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, New York, NY, USA, 17–19 October 2016.
- [28] Webb, J.; Hume, D. "Campus IoT collaboration and governance using the NIST cybersecurity framework". In *Proceedings of the Living in the Internet of Things: Cybersecurity of the IoT*, London, UK, 28–29 March 2018.
- [29] Pagliaro, F.; Mattoni, B.; Ponzo, G.; Corona, V.; Nardecchia, F.; Bisegna, F.; Gugliemetti, F. Sapienza "Smart campus: From the matrix approach to the applicative analysis of an optimized garbage collection system". In *Proceedings of the IEEE International Conference on Environment and Electrical Engineering and the IEEE Industrial and Commercial Power Systems*, Europe, Milan, Italy, 6–9 June 2017.
- [30] Verma, P.; Kumar, A.; Rathod, N.; Jain, P.; Mallikarjun, S.; Subramanian, R.; Amrutur, B.; Kumar, M.S.M.; Sundaresan, R. "Towards an IoT based water management system for a campus". In *Proceedings of the IEEE First International Smart Cities Conference (ISC2)*, Guadalajara, Mexico, 25–28 October 2015.
- [31] Lazaroiu, G.C.; Dumbrava, V.; Costoiu, M.; Teliceanu, M.; Roscia, M. "Energy-informatic-centric smart campus". In *Proceedings of the IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*, Florence, Italy, 7–10 June 2016.
- [32] Liang, Y.; Chen, Z. "Intelligent and Real-Time Data Acquisition for Medical Monitoring in Smart Campus". *IEEE Access* 6, 74836–74846, 2018.
- [33] Gjoreski, M.; Gjoreski, H.; Lutrek, M.; Gams, M. "Automatic Detection of Perceived Stress in Campus Students Using Smartphones". In *Proceedings of the International Conference on Intelligent Environments*, Prague, Czech Republic, 15-17 July 2015.
- [34] Ghirardini, Beatrice. "E-learning methodologies: A guide for designing and developing e-learning courses". Food and Agriculture Organization of the United Nations, 2011.
- [35] Smyrnowa-Trybulska, E. (Ed.). "E-learning and Intercultural Competences Development in Different Countries". Katowice-Cieszyn: Studio Noa for University of Silesia, 484 p. ISBN 978-83-60071-76-2, 2014.
- [36] Hwang, Gwo-Jen. "Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective." *Smart Learning Environments* 1.1, 4, 2014.
- [37] Kwok, L.M., Hui, Y.K. "The role of e-portfolio for Smart Life Long Learning" In: *Smart Universities, Smart Innovation, Systems and Technologies* 70, Springer International Publishing AG 2018, pp. 327-355, 2018.
- [38] Uskov V.L., Bakken J.P., Penumatsa A., Heinemann C., Rachakonda R. "Smart Pedagogy for Smart Universities". In: *Smart Education and e-Learning. SEEL 2017*. Smart Innovation, Systems and Technologies, vol 75. Springer, Cham, 2017.
- [39] Siemens, G., & Long, P. "Penetrating the fog: Analytics in learning and education". *EDUCAUSE review*, 46(5), 30, 2011.
- [40] Wagner, Ellen. and Phil Ice. "Data changes everything: Delivering on the promise of learning analytics in higher education." *Educause Review* 47.4 : 32, 2012.

- [41] Hrabowski, F. A. III & Suess, J., "Reclaiming the lead: higher education's future and implications for technology". *EDUCAUSE Review*, 45, 6 (November/December 2010). Retrieved October 30, 2014, from <http://www.educause.edu/library/ERM1068>
- [42] Tulasi. B. "Significance of big data and analytics in higher education." *International Journal of Computer Applications* 68.14, 2013.
- [43] Huang Ronghuai, Hu Yongbin, Yang Junfeng, and Xiao Guangde, "The functions of smart classroom in smart learning age," *Open Education Research*.;18(2):22-27, 2012.
- [44] Zhu, Z., Yu, M. & Riezebos, P. "A research framework of smart education". *Smart Learning Environments*. 3, 4 , 2016.
- [45] James Ransom, "Smart Places: How universities are shaping a new wave of smart cities", *British Council journal*, 2019.
- [46] Bhatia, Satinder, "Sustainable Smart Universities for Smart Cities". *Journal of Economics, Management and Trade*. 21. 1-11. 10.9734/JEMT/2018/44521, 2018.
- [47] Living Smart Campus programme, University of Twente, 2016. <https://www.utwente.nl/en/organisation/news-agenda/special/2016/living-smart-campus>
- [48] Chris Pearce, Smart Campus - Digital Master plan. 2019. <https://www.gla.ac.uk/myglasgow/worldchangingglasgow/projects/smartcampusproject/>
- [49] Huy Quang Tran, Nhat Tien Le, Quang Luong Nguyen, Dat Tan La and Thu Thi Anh Nguyen. "An effective model for integrated face detection and recognition". The 8th Conference on Information Technology and Its Applicatio. 2019;. p.23-30.



# Characteristics of Recycled Reinforced Concrete at High Temperatures

Nguyen Thanh Hung

Department of Civil Engineering, Ho  
Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
hungnt@hcmute.edu.vn

Dao Duy Kien\*

Department of Civil Engineering, Ho  
Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
kiendd@hcmute.edu.vn

Nguyen Van Khoa

Department of Civil Engineering, Ho  
Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
khoanv@hcmute.edu.vn

Tran Minh Hieu

Department of Civil Engineering,  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
16149044@student.hcmute.edu.vn

Doan Dinh Thien Vuong

Department of Civil Engineering,  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
doandinhthienvuongcons@gmail.com

**Abstract**— This study investigates the behavior of reinforced concrete and recycled fibre reinforced concrete influenced to temperatures of 200 ° C - 500 ° C, on two types of concrete such as recycled concrete, recycled fibre concrete. The experiment was conducted at the Structural Engineering laboratory at Ho Chi Minh City University of Technical Education. The characteristics, those are the failure, compressive strength, cracking were analyzed. Based on that, compressive strength of recycled fibre concrete has clearly changed in comparison with recycled concrete when subjected to temperatures of 200°C - 500°C. On the other hand, failure mechanism and shape of cracking are to-tally difference between two types of specimens.

**Keywords**— Recycled concrete, Compressive strength, high temperature.

## I. INTRODUCTION

Overseas, current studies have addressed several issues, such as: The initial study on the structural performance of RCA has been published in Japan [1]. The authors Luis Evangelista and Jorge de Brito [2] have studied the use of recycled reinforced concrete to replace fine natural aggregate (sand) without jeopardizing the mechanical properties of concrete and for Replacement rate up to 30%. Authors Luis Evan-gelista, Jorge de Brito and Pereira [3] studied the effects of superplasticizers made from recycled concrete. Bai and Sun (Bai WH, et al., 2010) [4] used 8 concrete beams from 8 to 10 years to make concrete scrap aggregates. Model of concrete beams with the replacement of concrete fragments with different levels of 50%, 70%, and 100%. They observed a similar crack pattern, but the deflection and crack width increased with the increase of the replacement rate of concrete. They also concluded that the replacement rate of concrete fragments did not significantly affect the maximum crack of reinforced concrete beams: Authors T. Li, J. Xiao, C. Zhu, and Z. Zhong (Construction and Building Materials, vol. 120, 2016) [5] studied a new large-sized aggregate, large-sized recycled aggregate (LRCA) with the most significant edge of 80mm. It has been studied in order to simplify the grinding process of concrete to produce highly efficient

replacement aggregate. The block compression test showed that the strength of LRCA concrete compared to NAC concrete difference is relatively small. When the LRCA size is 80 mm, and the proportion of replacement components accounts for 40%, the compressive strength only decreases by about 14%. When the component ratio is lower than 30%, the compressive strength is not significantly reduced. The compression ratio along LRCA concrete is 12% lower than NAC concrete. Tensile strength is lower than 10%. The results show that LRCA has excellent mechanical properties and can be used in concrete such as pile foundations and foundation supports.

In the country, recently, the authors of Le Anh Thang, Nguyen Thanh Hung and Phan Cong Vu Duc have studied the behavior of reinforced concrete beams with recycled concrete components [6].

The above studies have contributed to the technical improvement of the proper-ties of RAC concrete to apply in practice. In addition to the research on the mechanical properties and durability of RAC and the replacement percentage of RCA aggregate, it is necessary to pay attention to the heat resistance and workability of concrete after exposure to high temperatures. When concrete is exposed to high temperatures, temperatures can cause bond destruction, the mechanical ability of concrete, durability. Several studies on concrete limits have shown that after exposure to high temperatures, the intensity of RAC varies significantly, with the higher the temperature, the longer the exposure time, the higher the decrease in concrete strength. The study also shows that the remaining strength of RAC concrete can be superior or equivalent to that of ordinary concrete. The two main factors causing the decrease in concrete strength are: 1. The increase in steam pressure arising from moisture evaporation; 2. The formation and development of cracks due to differential thermal stresses between different parts of concrete (eg surface and core). Because the moisture content and aggregates differ between RAC and NC, the effects of these two factors will be different for RAC and NC.

Based on the above issues, this paper delves into RAC and NC when exposed to high temperatures, including testing on compressive properties, residual strength, elastic modulus, and relationships, concrete stress. This study aims to find effective measures to reduce the adverse effects of temperature on RAC and NC concrete. The results presented in this valuable article help to understand more about the RAC and NC in heat resistance. The study can refer to the application of RAC techniques to the structures and structures of buildings.

## II. EXPERIMENT PROCEDURES

### A. Test specimens

The test specimens were conducted on a recycled concrete model(RAC). Recycled concrete (RAC) is a concrete with a recycled concrete aggregate component replacing natural stone aggregate. In order to conduct experiments, it is necessary to cast cylindrical samples with standard dimensions of 100mm × 200mm (diameter × height) as shown in Fig. 1.



Figure 1. Specimens

### B. Test procedures and measurement

After 28 days of specimens are stored, preheated samples need to be dried in an oven at 80°C for 6 hours.

After completion of the drying period, the samples were arranged in rows, one row as a group. Each group was heated to the Naberthem Chamber Furnaces LH15 / 14 at the

laboratory with maximum temperature is 3000°C. This is a kiln with a quite large capacity, so it is possible to reach a heating speed of up to 8°C / min and this is also the maximum heating rate that can be achieved. Set up the heating program by turning the control knob located on the control box of the oven and adjusting the set parameters. With a specialized furnace with a 5-sided chamber structure with twisted steel wires at the start of the program, these wires will get hotter with each temperature level until the program's set temperature is reached. The structure of the closed chamber of the furnace, during the heating process ensures uniformity and uniformity.

The specimens were heated to discuss the effect of temperature at 0, 200, 300 and 400°C on the compressive strength

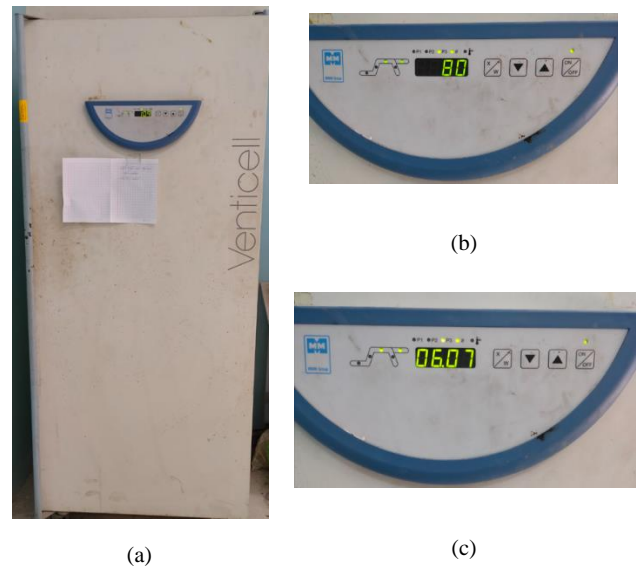


Figure 2. Drying machine for specimens before heating. (a) Venticell drying oven, (b) The temperature class used for drying the test samples was 80 °C, (c) The test samples were dried for 6 hours.



Figure 3. Tensile/compression machine and Servo-plus evolution touch screen system

### III. RESULTS AND DISCUSSION

The compression test results, the intensity results, the elastic module are processed and collected by the touch screen system, and the graph of load versus time.

TABLE 1. SUMMARY OF RECYCLED CONCRETE DATA (RAC) BY TEMPERATURE LEVEL

Temperature		200°C	300°C	400°C
Max Load	145.002	96.221	142.54	151.97
Area (mm <sup>2</sup> )	7853.98	7853.98	7853.9	7853.9
Max strength (Mpa)	18.462	12.251	18.150	19.350
Initial length (mm)	201.5	200	201.5	199
Deformation (mm)	1.5	2.5	3	5
Relative deformation	0.0075	0.0125	0.0099	0.0251
Deformation modulus (Mpa)	2.462 x10 <sup>3</sup>	0.98 x10 <sup>3</sup>	1.829 x10 <sup>3</sup>	0.77x10 <sup>3</sup>

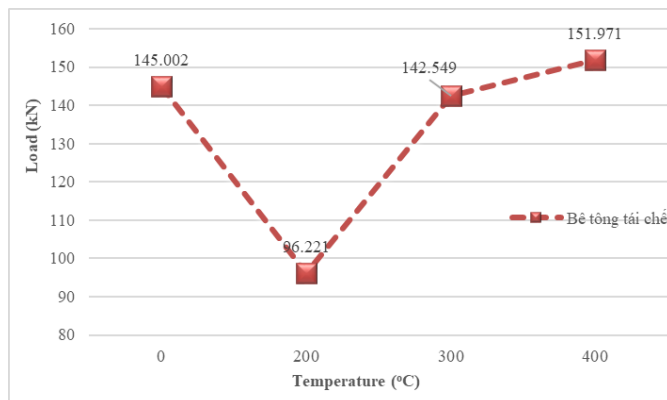


Figure 4. Graph comparing the strength of concrete at different temperatures

Based on Figure 4, we see that:

- Temperature 2000C compressive strength of recycled concrete decreased by 33.64% compared to non-heated NC.
- Temperature 3000C compressive strength of recycled concrete decreased by 1.69% compared to non-heated NC.
- Temperature 4000C compressive strength of recycled concrete increased by 4.81% compared to non-heated NC.

Because RAC uses RCA aggregates in mortar and concrete to increase shrinkage be-cause recycled concrete consumes substantial amounts of water, the heating process leads to the expansion of pressure of pore water particles to

form the cracks and destruction of the link between cement and aggregate cause a sudden decrease in intensity (at 2000C). Then, when the temperature of concrete strength increases again due to the disappearance of the water pore particles inside the concrete and the chemical transformation recombines between cement and aggregate, the concrete strength will continue. Increase until the limit is reached and may decrease after that.

### IV. CONCLUSIONS

Experimental research project on the effect of temperature on the strength of conventional and recycled concrete has drawn some conclusions:

1. As concrete changes color gradually with increasing temperature, the darker the temperature of the concrete, the more it proves that the temperature affects the properties and produces a reaction during the concrete heating process.

2. The compressive strength of concrete is also affected when exposed to high temperatures:

- For RAC, with functional water absorption capacity, the intensity has a sudden decline of 60% of the RAC intensity (at 2000C), and at other temperatures, the aver-age is kept at 100% RAC intensity.

3. The change in intensity also entails a change in the elastic module in the opposite direction. Results showed that when the strength of concrete increased, the elastic module decreased compared to the concrete strength. This shows that the temperature has influenced the destruction of concrete before and during heating.

### REFERENCES

- [1] Yagishita F, Sano M and Yamada M, "Behavior of reinforced concrete beams containing recycled coarse aggregate, Demolition and reuse of concrete & masonry RILEM proceeding 23". - [s.l.]: Demolition and reuse of concrete & masonry RILEM proceeding 23, 1994. - pp. 331–342.
- [2] Luis Evangelista and Jorge de Brito (2007), "Mechanical behaviour of concrete made with fine recycled concrete aggregates", *Cement and Concrete Composites* 29(5). - 2007. - pp. 397-401.
- [3] Luis Evangelista, Jorge de Brito and Pereira , "The effect of superplasticisers on the workability and compressive strength of concrete made with fine recycled concrete aggregates", *Instituto Superior de Engenharia de Lisboa, R. Conselheiro Emídio Navarro*. - pp. 1959-1001.
- [4] Bai WH and Sun BX, "Experimental study on flexural behavior of recycled coarse aggregate concrete beam [Journal]" // *Applied Mechanics and Materials* 29. - 2010. - pp. 543-548
- [5] T. Li, J. Xiao, C. Zhu, và Z. Zhong, " Experimental study on mechanical behaviors of concrete with large-size recycled coarse aggregate", *Construction and Building Materials*, vol. 120, 2016- pp.321 -328 )
- [6] Anh Thang Le, Thanh Hung Nguyen, Cong Vu Duc Phan (2019), " A Study on Behavior of Reinforcement Concrete Beam using the Recycled Concrete", *Proceedings of the 5th International Conference on Geotechnics, Civil Engineering Works and Structures 2019* -pp. 379-384.
- [7] Nguyễn Thanh Hưng, Lê Anh Thắng, Lê Ngọc Phương Thanh Thực nghiệm cường độ chịu nén của bê tông có thành phần cốt liệu là bê tông tái chế [Journal] / Tạp chí xây dựng Việt Nam, số 07. - 2018. - pp. 34-38.

# The Influence of Raft Thickness on the Behaviour of Piled Raft Foundation

Tong Nguyen  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
tongn@hcmute.edu.vn

Phuong Le  
HCMC University of Technology and  
Education  
Ho Chi Minh City, Vietnam  
phuongle@hcmute.edu.vn

Viet Tran  
Optimum construction Consulting Co.,  
LTD (Opticons)  
Ho Chi Minh City, Vietnam  
trannguyenviet87@gmail.com

**Abstract**— The behaviour of piled raft foundation is influenced by several factors such as the thickness of raft, the number of piles, the length of piles as well as the types of soil in which the foundation is inserted. The piled raft foundation in the article is derived from the Vietcombank Tower's foundation and is analyzed by CSI SAFE software. This numerical modelling is verified by comparing the monitoring result of settlement. Based on this model, the authors performed the different analyzes of the piled raft foundation by varying the amount of piles and the thickness of raft with the same condition of the ground and pile length. The results demonstrate that the choice of raft thickness is the main factor that affects the optimization of the piled raft foundation.

**Keywords**—pile raft, numerical modelling, raft thickness

## I. INTRODUCTION

The piled raft foundation consists of three load-bearing elements: piles, raft, and subsoil. According to their stiffness, the raft distributes the total load transfer from the structure as contact pressure below the raft and load over each of the piles. In the conventional design, either the raft or the piles is designed to support the building load with adequate safety against bearing capacity failure and against loss of overall stability. In the piled raft foundation, the contributions of the raft and piles are taken into consideration to verify the ultimate bearing capacity and serviceability of the overall system.

Many studies of the analyzing piled raft foundation are published. The approaches can be divided into analytical methods and numerical methods such as finite element methods, boundary element methods, or hybrid methods. Randolph [1] was suggested new analytical approaches for the design of pile groups and piled raft foundations to focus on the settlement issue rather than the capacity. Russo [2] presented an approximated numerical methods for analysis for the analysis of piled raft foundation, in which the raft is modeled as a thin plate and the piles as interacting non-linear springs. Both the raft and the piles are interacting with the soil which is modeled as an elastic layer. Prakoso and Kulhawy [3] analyzed the piled raft foundation using a simplified linear elastic and non-linear plane strain finite element models.

Based on the above approaches, several studies on the influence of factors such as raft thickness, pile distance, pile length, pile layout and etc... on piled raft foundation behavior was also published a lot. Prakoso and Kulhawy [3] demonstrated that the raft and pile group system geometries and pile compression capacity have an effect on the average and differential displacements, raft bending moments, and pile

butt load ratio. Poulos [4] suggested three different stages of design for piled raft foundation. In the first stage, the effect of the number of piles on load capacity and settlement are assessed through an approximate analysis. The second stage is a more detailed examination to assess where piles are required. The third is a detailed design phase in which a more refined analysis is employed to confirm the optimum number and local of the piles. El-Garhy, Galil, Youssef, & Raia[5] showed that the raft thickness has an important effect on the differential settlement but has a negligible effect on the average settlement and load distribution between piles and the raft. Tang, Pei & Zhao [6] noted that for a spacing greater than five times the diameter of the pile, the raft and the pile behavior independently and piles may reach their ultimate load capacity.

It is clear that there are many factors that influence the optimization of the piled raft foundation. The above approaches have been well used but still complicated, especially when used directly for design work. The authors proposed a simpler approach as mentioned in [7]. In this article, the authors used that approach to study the effect of raft thickness on the behaviour of the piled raft foundation of Vietcombank Tower. Based on the back-calculated model, the authors rebuilt the different analyzed models by varying the amount of piles and the thickness of raft without changing the ground condition and the pile length. With some results obtained, the article helps engineers better judge in their design work when designing the piled raft foundation.

## II. MODELLING OF PILED RAFT FOUNDATION

### A. Vietcombank Tower overview

The building is located at No. 5, Me Linh Square, District 1, Ho Chi Minh City. It consists of 4 basements and 40 floors, the total height of the building is 206 meters, including the roof and antenna tower.

The raft structure is made of B25 graded concrete and is 2.75 meters thick (tower area), 1.00 meters thick (podium area). The depth of foundation bottom from ground level is 15.55 meters and 13.48 meters respectively. Therefore, the raft foundation is entirely placed on layer 4 (silty clayey sand, greyish-yellow, loose to dense. Layer bottom of 34.5-36.7 meters and thickness of 22.6-27.7 meters).

### B. Analytical model of piled raft foundation

The raft is modeled as a thin plate, the piles as point springs and soil as soil subgrade. The retaining structure around the basement of the building by diaphragm wall is modeled as line spring. CSI SAFE software [8] is used to analyze this model. The method of estimating the parameters

of point spring, line spring and soil subgrade was presented in [7]. In the article, the authors only present the results obtained as follow:

- Soil Subgrade: 4540 kN/m<sup>3</sup>.
- Point Springs: These values depend on the location of the piles. For the corner piles, their value is 815 kN/mm, while for the edge piles it is 844 kN/mm. At positions ¼ width of the raft, their value is 768 kN/m. And in the central area of the raft, it is 752 kN/m.
- Line Spring: 357.5 kN/mm.

The design capacity of the pile is 24700 kN.

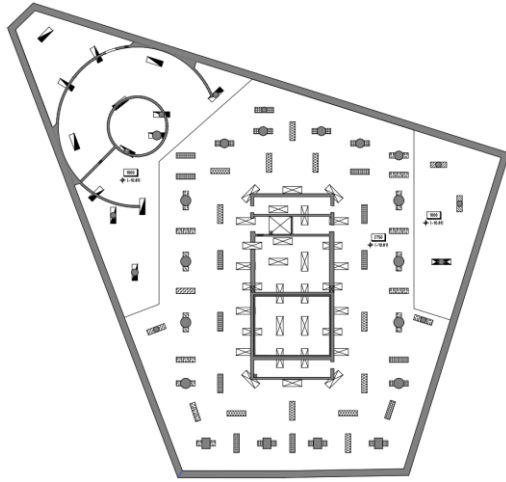


Fig. 1. Foundation plan and pile configuration.

The construction process carried out the settlement monitoring of piled raft foundation. There are 16 points marked from S1 to S16. This settlement monitoring took place from 30/07/2012 to 16/05/2014. The number of cycles to be performed is 30 cycles. The layout of settlement monitoring points is shown in Fig. 2

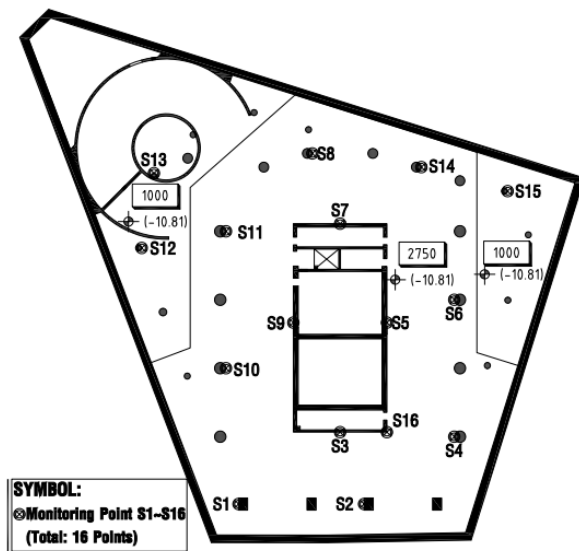


Fig. 2. Layout of settlement monitoring points.

### C. Verification of the analytical model

The displacement of the piled raft foundation from the analytical model by using CSI SAFE software is compared with the settlement monitoring results. These are shown in Table I and Fig. 3.

TABLE I. VERTICAL DISPLACEMENT OF PILED RAFT FOUNDATION BY CSI SAFE AND BY MONITORING

Point	Vertical displacement of piled raft foundation		
	Monitoring	CSI SAFE	Error (%)
S1	12.56	8.27	-34.2
S2	10.36	11.80	13.9
S3	19.02	19.56	2.8
S4	13.22	12.47	-5.7
S5	20.95	24.47	16.8
S6	11.77	18.56	57.7
S7	20.40	20.27	-0.6
S8	14.31	12.91	-9.8
S9	19.91	22.43	12.7
S10	13.00	15.77	21.3
S11	18.27	15.87	-13.1
S12	7.24	7.57	4.6
S13	10.95	10.35	-5.5
S14	13.05	11.75	-10.0
S15	11.03	8.40	-23.8
S16	22.11	20.35	-8.0

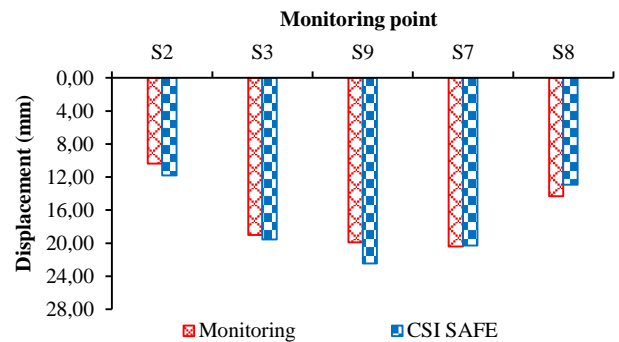


Fig. 3. Displacement of points along the centreline by CSI SAFE and by monitoring

According to the results shown in Table I, the vertical displacement of the piled raft foundation in the central area (including S3, S5, S7, S9, and S16) have higher values than those in other positions. The displacement errors at these points in this area range from -8% (S16) to +16.8% (S5) if compared with the results of settlement monitoring in similar positions. The errors in other positions are much higher (with an average error of  $\pm 25\%$ ), especially in the edge positions (including S1, S2, S4, S6, S8, S10, S11, and S14). The average error of the model is approximately  $\pm 20\%$ . The displacement of points through a section (including S2, S3, S9, S7, and S8) shown in Fig. 3 indicates that the displacement configuration from the results by monitoring and by CSI SAFE is similar. It means that the proposed analytical model described the piled raft foundation system reliably. Therefore, this model can be



used to investigate some factors that affect the behavior of the piled raft foundation.

### III. PARAMETRIC STUDY

The parametric study was performed to link and establish the influence of raft thickness and the number of piles on the behavior of the piled raft foundation.

Based on the above analytical model, the authors added 14 points denoted from M1 to M14 to analyzed the average settlement and differential settlement of piled raft foundation. The layout of settlement monitoring points and additional reference points is shown in Fig. 4.

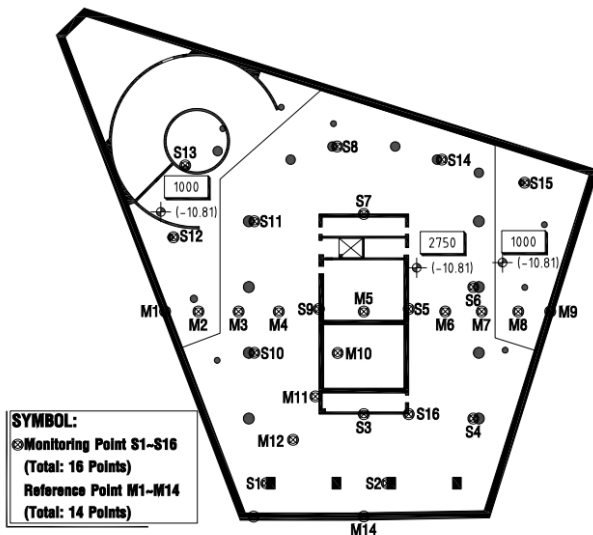


Fig. 4. Layout of settlement monitoring points and additional reference points.

Raft thickness was changed from 1.00 meter to 4.00 meters (1.00m, 1.50m, 2.00m, 2.50m, 2.75m, 3.00m, 3.50, and 4.00m). The number of piles was changed from 98 piles to 73 piles (98 piles, 86 piles, 79 piles, and 73 piles), with the pile to raft area ratio 8.9%, 7.8%, 7.1%, and 6.6% respectively. Thus, there are 32 analysis models to evaluate some criteria in the design of the pile-raft foundation including the average settlement, the differential settlement, the raft bending moment, the surface pressure of the raft bottom, and the pile reaction. The analysis models are presented in Table II where MH1-5 is the initial analytical model.

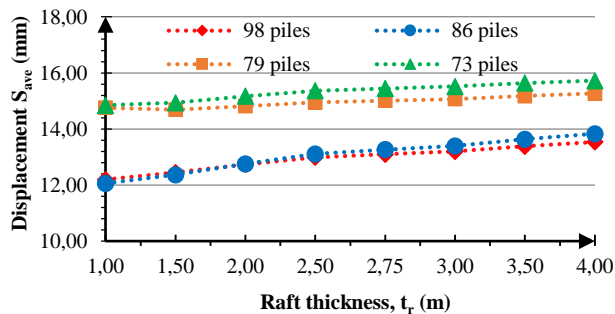


Fig. 5. Layout of settlement monitoring points and additional reference points.

TABLE II. ANALYTICAL MODELS WITH DIFFERENT RAFT THICKNESS AND NUMBER OF PILES

Model	Raft thickness (m)	Number of piles	Ratio $R_{area}$ (%)	Model	Raft thickness (m)	Number of piles	Ratio $R_{area}$ (%)
MH1-1	1.00	98	8.9	MH3-1	1.00	79	7.1
MH1-2	1.50			MH3-2	1.50		
MH1-3	2.00			MH3-3	2.00		
MH1-4	2.50			MH3-4	2.50		
MH1-5	2.75			MH3-5	2.75		
MH1-6	3.00			MH3-6	3.00		
MH1-7	3.50			MH3-7	3.50		
MH1-8	4.00			MH3-8	4.00		
MH2-1	1.00	86	7.8	MH4-1	1.00	73	6.6
MH2-2	1.50			MH4-2	1.50		
MH2-3	2.00			MH4-3	2.00		
MH2-4	2.50			MH4-4	2.50		
MH2-5	2.75			MH4-5	2.75		
MH2-6	3.00			MH4-6	3.00		
MH2-7	3.50			MH4-7	3.50		
MH2-8	4.00			MH4-8	4.00		

#### A. Average settlement of piled raft foundation

As the raft thickness changes from 2.5 meters to 4 meters, the average settlement of piled raft foundation does not change significantly (Fig. 5) if the number of piles is constant. But when the raft thickness is less than 2.5 meters, the average settlement of piled raft foundation increases linearly regardless of the number of piles. Perhaps, the raft is not stiff enough to resist the load from the superstructure. As a result, load distribution is limited locally in high load areas. According to Fig. 6 the displacement at points M4, M5, S5, and S6 increases or decreases dramatically when the raft thickness is less than 2.5 meters, and these points are located in the elevator area where the load is high. It is obvious that if the raft has a suitable stiffness, the raft thickness has a negligible effect on the average settlement. This evidence is compatible with the results obtained in ElGarhy et al [5].

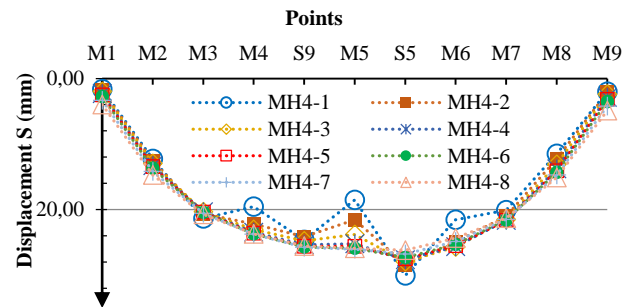


Fig. 6. Displacement of points along the centreline (MH4 models)

### B. Differential settlement of piled raft foundation

TABLE III. THE DISPLACEMENT OF POINTS AND THE DIFFERENTIAL SETTLEMENT ALONG THE CENTRELINE (MH4 MODELS)

Points	MH4-1	MH4-2	MH4-3	MH4-4
M1	1.65	1.79	1.98	2.3
M2	12.27	12.64	12.95	13.28
M3	21.34	20.61	20.34	20.31
M4	19.53	22.17	23.02	23.47
S9	24.64	24.14	24.76	25.38
M5	18.52	21.55	23.79	25.19
S5	30.02	28.4	27.96	27.67
M6	21.55	24.96	25.68	25.63
M7	20.08	21	21.51	21.66
M8	11.51	12.28	13.13	13.79
M9	2.02	2.14	2.44	2.92
$\Delta S_{(S5-M1)}$	28.37	26.61	25.98	25.37
$\Delta S_{(S5-M9)}$	28	26.26	25.52	24.75
Points	MH4-5	MH4-6	MH4-7	MH4-8
M1	2.51	2.75	3.32	3.96
M2	13.47	13.7	14.2	14.74
M3	20.36	20.42	20.58	20.74
M4	23.61	23.7	23.76	23.69
S9	25.58	25.71	25.74	25.54
M5	25.61	25.88	26.05	25.88
S5	27.49	27.29	26.78	26.17
M6	25.49	25.3	24.85	24.35
M7	21.65	21.61	21.48	21.31
M8	14.06	14.31	14.73	15.1
M9	3.22	3.54	4.22	4.91
$\Delta S_{(S5-M1)}$	24.98	24.54	23.46	22.21
$\Delta S_{(S5-M9)}$	24.27	23.75	22.56	21.26

Table III shows that when the thickness of raft increases, the displacement on the raft edge (point M1 and M9) increases and the displacement in the center of the raft (point S9, M5, and S5) decreases, resulting in a decrease in the difference in the settlement. As the thickness of raft changes from 2.5 meters (MH4-4) to 4 meters (M4-8), the displacement at point M1 increases from 2.3 mm to 3.96 mm (an increase of 75%), while the displacement at point S5 decreases from 27.67 mm to 26.17 mm (a decrease of 5%). However, the average error of the analytical model is  $\pm 10\%$  in the central area of the raft and  $\pm 25\%$  in other areas of the raft. So the displacement of points located in the central area of the raft almost does not change, but the displacement of points located in the raft edge increases. Thus, the differential settlement tends to decrease (from 25.37 mm to 22.21 mm) with increasing the raft thickness (from 2.5 meters to 4 meters). Although Fig. 7c demonstrates that this statement is right, Fig 7a and Fig 7b show that the differential settlement of the raft between the corner point and the edge point or between the corner point

and the central point is almost identical to a greater 2.5 meters thick raft. According to Fig. 7, as the number of piles increases from 73 piles to 98 piles, the differential settlement decreases but is constant if the number of piles increased from 86 to 98 piles for all sections. It is clear that as the raft thickness increases to a certain extent, the effect on the differential settlement is negligible. Note with a raft thickness of fewer than 2.5 meters, load distribution is local at several locations, the evaluation of the differential settlement with the approach mentioned in Table III is not reliable.

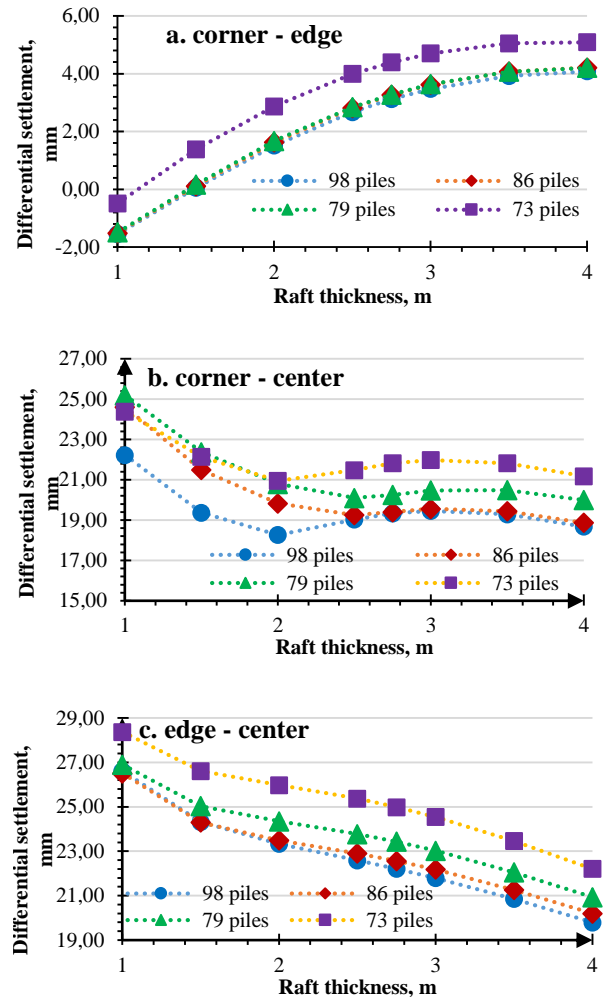


Fig. 7. Differential settlement along several sections

### C. Raft bending moment

Fig. 8 indicates that as the raft thickness increases, the maximum bending moment of the raft increases. As the raft thickness increases from 2.5 meters to 4.0 meters, the maximum bending moment of the raft increases by approximately 35%. The analytical models have an average error of  $\pm 20\%$ . Thus, the magnitude of maximum moments represents the uptrend of the maximum bending moment with increasing the raft thickness.

The ultimate bending moment depends mainly on the strength of concrete, the strength of reinforcement, and the raft thickness. As the raft thickness is increases, the ultimate bending moment also increases. Fig. 8 shows that as the raft thickness is greater than 3.0 meters, the ultimate bending moment starts to increase dramatically and its magnitude is much greater than the maximum bending moment of the raft.

While the raft thickness is approximately 2.5 meters to 3.0 meters, the difference between the ultimate bending moment and maximum bending moment is quite small. If the raft thickness is less than 2.5 meters, the raft is not capable of bearing.

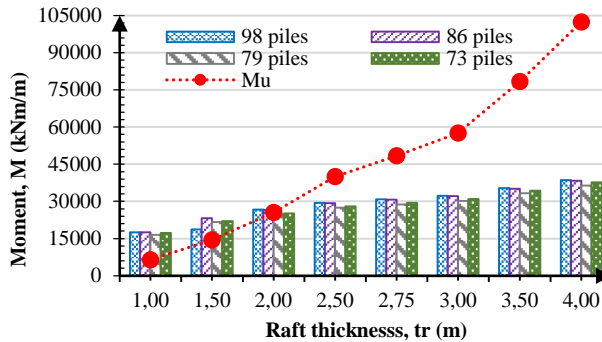


Fig. 8. Maximum bending moments and ultimate bending moments of the raft ( $M_u$ )

#### D. Surface pressure of raft bottom

Fig. 9 indicates that with the same number of piles if the raft thickness increases, the surface pressure at the raft bottom decreases. As the number of piles decreases, the surface pressure at the raft bottom increases. According to Fig. 9, as the raft thickness is greater than 2.5 meters, the surface pressure at the raft bottom decreases linearly and this trend is independent on the number of piles. As the number of piles increases to a certain extent, the surface pressure at the raft bottom reaches asymptotically a limit for any raft thickness.

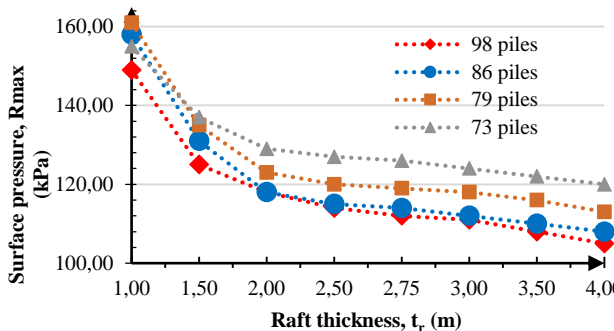


Fig. 9. The surface pressure of raft bottom

#### E. Pile reaction

Fig. 10 shows that as the raft thickness increases, the maximum pile reaction at the central area and the corner area decreases, while the maximum pile reaction at the edge area increase. Note as the raft thickness is less than 2.0 meters, the pile reaction at the central area is much greater than the design capacity of the pile, while the pile reaction at other areas is much smaller than the design capacity of the pile. As the raft thickness increases from 2.5 meters to 4.0 meters, the pile reaction at the central area tends to decrease, but the pile reaction at the corner area or the edge area is almost constant. This trend occurs regardless of the number of piles. There is a relationship between the flexibility of the raft and the redistribution of the pile reaction. However, as the raft thickness increases to a certain extent, the redistribution of the pile reaction depends only on the tributary load area without the raft thickness.

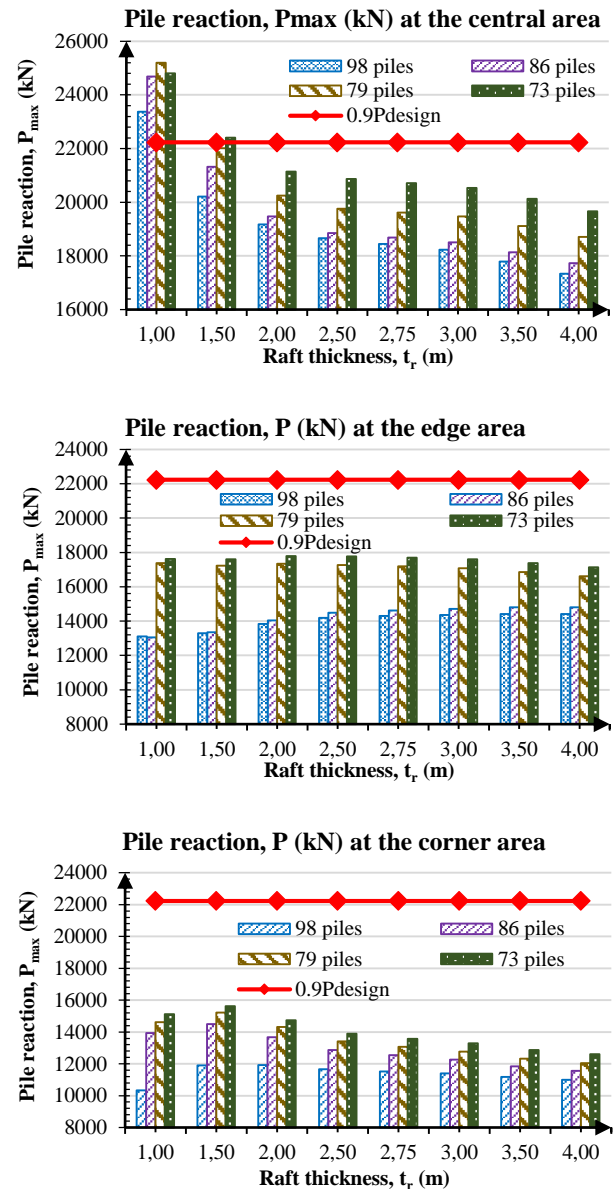


Fig. 10. Pile reaction at different positions ( $P_{design}$  – The design capacity of the piles)

#### F. Load sharing between piles and raft

Fig. 11 indicates that as the raft thickness increases, the load sharing between the piles and the raft does not almost change. it only depends on the number of piles. the load sharing coefficient of piles decreases with the number of piles. as the raft thickness is small, the load distribution may be ununiform between locations under the raft. even at high load locations, the pile cannot be sufficient to resist the load. but the total load sharing coefficient of the pile is constant regardless of the raft thickness.

#### IV. CONCLUSION

Raft thickness does not have any appreciable effect on the average settlement of the piled raft foundation. However, the differential settlement of the piled raft foundation increases with raft thickness up to a limit, and beyond it, raft thickness has a negligible effect on its differential settlement. At a certain extent of raft thickness, the average settlement and the differential settlement decrease with an increase in the number of piles, but it doesn't decrease more if the number of piles

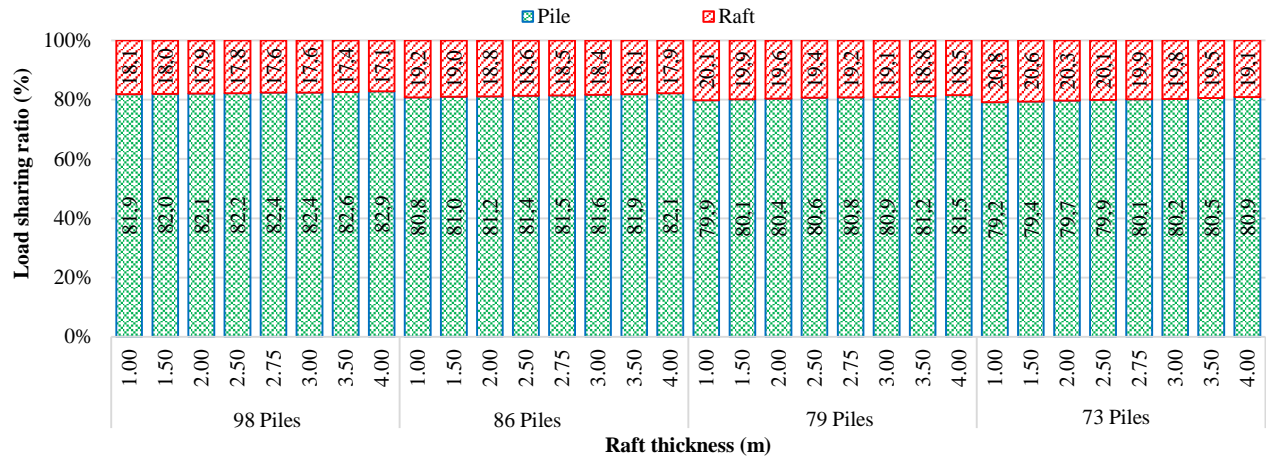


Fig. 11. Load sharing between raft and piles

continues increasing. It means that there is a limit to the number of piles and the raft thickness in reducing settlement. Besides, the minimum raft thickness is also necessary to avoid the dramatic displacement of the raft at high load locations.

The maximum bending moment of raft increases with raft thickness. However, there is a limit to raft thickness to match the ultimate bending moment for optimum design.

The surface pressure at the raft bottom decreases if raft thickness increases. As the number of piles decreases, the surface pressure increases. The pressure surface reaches asymptotically a limit for any raft thickness. However, the minimum raft thickness is necessary to increase the contact between raft and soil.

As the raft thickness increases, the load redistribution from the superstructure to piles is more uniform at locations. The maximum pile reactions at the central area tend to reduce, while the maximum pile reactions at other areas tend to increase. However, as the raft thickness increases to a certain extent, the redistribution of the pile reaction doesn't depend on the raft thickness.

The load sharing between the piles and the raft is independent on the raft thickness. The number of piles is the main factor affecting the load sharing foundation.

## REFERENCES

- [1] M. F. Randolph, "Design methods for pile groups and piled rafts," in *13th International Conference for Soil Mechanics and Foundation Engineering*, New Delhi, 1994.
- [2] G. Russo, "Numerical analysis of piled rafts," *International Journal for Numerical and Analytical Methods in Geomechanics*, pp. 477-493, 1998.
- [3] Prakoso, W. A. and Kulhawy, F.H, "Contribution to piled raft foundation design," *Journal of Geotechnical and Geoenvironmental Engineering*, ASCE, pp. 17-24, 2001.
- [4] H. G. Poulos, "Piled raft foundation: design and applications.," *Geotechnique*, pp. 95-113, 2001.
- [5] El-Garhy, B., Galil, A. A., Youssef, A.-F., & Raia, M. A., "Behavior of raft on settlement reducing piles: Experimental model study," *Journal of Rock Mechanics and Geotechnical Engineering*, pp. 389-399, 2013.
- [6] Tang, Y. J., Pei, J., & Zhao, X. H, "Design and measurement of piled-raft foundations," *Proceedings of the Institution of Civil Engineers - Geotechnical Engineering*, pp. 461-475, 2014.
- [7] Viet Tran, Phuong Le, Tong Nguyen, Hung Nguyen Si, "A new approach for the design of piled raft foundations using CSI SAFE software," *Vietnam Journal of Construction - Copyright Vietnam Ministry of Construction*, pp. 72-79, 2018.
- [8] Computer and Structures, Inc, "CSI analysis reference manual for SAP2000, ETABS, SAFE," Computer and Structures, Inc, 1995 University Avenue Berkeley, California 94704 USA, 2018.

# Parallel Multi-Population Technique for Meta-Heuristic Algorithms on Multi Core Processor

Nguyen Tien Dat

Faculty of Electrical &  
Electronics Engineering  
Ho Chi Minh University of  
Technology, VNU-HCM  
Ho Chi Minh City, Viet Nam  
ntdat@hcmut.edu.vn

Cao Van Kien

Faculty of Electrical &  
Electronics Engineering  
Industrial University of Ho Chi  
Minh City  
Ho Chi Minh City, Viet Nam  
caovankien@iuh.edu.vn

Ho Pham Huy Anh

Faculty of Electrical &  
Electronics Engineering  
Ho Chi Minh University of  
Technology, VNU-HCM  
Ho Chi Minh City, Viet Nam  
hphanh@hcmut.edu.vn

Nguyen Ngoc Son

Faculty of Electrical &  
Electronics Engineering  
Industrial University of Ho Chi  
Minh City  
Ho Chi Minh City, Viet Nam  
nguyenngocson@iuh.edu.vn

**Abstract**—This paper investigates parallelization method for meta-heuristic algorithms such as Particle Swarm Optimization (PSO) and Differential Evolution (DE) on multi-core processor to reach eventually fast execution and stable result. In PSO or DE algorithm, all of member in initial population are created to search the best place in which the value of member in that place is satisfied the output criteria. As the parallelization method, the searching region is separated into many sub-regions which are executed with optimized algorithm on multi-core processor. The structure of meta-heuristic algorithms is rebuilt to execute in parallel multi population mode. The benchmark functions such as Rosenbrock, Griewank, Ackley and Michalewicz are used to test those proposed algorithms. The results show that the proposed parallel multi-population technique applied on PSO and DE algorithm has a competitive performance compared to the standard ones. The parallel multi-population technique shows better result which proves more precise and stable. Especially, meta-heuristic algorithms running in parallel multi-population mode execute quite convincingly faster than standard ones.

**Keywords**—Particle Swarm Optimization (PSO) - Differential Evolution (DE) - Parallel technique - Multi-strategy

## I. INTRODUCTION

The Particle Swarm Optimization (PSO) algorithm was developed by Kennedy and Eberhart [1,2] and Differential Evolution (DE) was proposed by Storn and Price [3], which are powerful basic algorithms that have been used for solving different types of optimization problems. The principle of PSO and DE is based on the hypothesis that social sharing of information among specifications aim to reach an evolutionary advantage. Both of those algorithms have been widely used to searching for best stage in a specified range in which random population is created.

Recently, many research related to PSO algorithms which is applied in many areas such as optimal problem [4], modification of landslide susceptibility mapping [5], estimating  $\alpha$  ratio in driven piles [6], path planning [7], color image segmentation [8]. And amount of research using DE algorithms to seek for the best solution such as minimizing paths on surfaces [9], determination of nano-aerosol Size Distribution [10], single objective optimization problems [11], multi-modal multi-objective optimization [12]. Parallel computing is the simultaneous use of multiple computing resources to solve a computational problem by breaking it into discrete parts. This computation process on a single machine as well as on multiple machines. Single machine processing includes computers utilizing multi-core, multiprocessor, and GPU with multiple processing elements.

Hung and Wang [17] proposed GPU-accelerated PSO (GPSO) by implementing a thread pool model with GPSO on a GPU, aimed to accelerate PSO search operation for higher dimension problems with large number of particles.

With above referred application of PSO and DE algorithms, this paper propose the method that splitting searching region into pieces of sub-region, in which the independence sub-population is created on each region, aimed to greatly decrease computing time base on utilizing the power of multi-core processor. The structure of PSO and DE are changed to split up and all of sub-regions are simultaneously searching by sub-population. After all, this technique is returned a best vector that contains global best stage of each sub-population. The global best stage of main-population is chosen base on it. In recent years, new types of hardware that deliver massive amounts of parallel processing power so running parallel multi population is possible. Moreover, recent CPU architectures have significant modifications leading to high-performance computing capabilities. One of those high-performance computing capabilities is Parallel Computing Toolbox that can solve intensive computing and data problems using multi-core processors. High-level constructions, such as parallel for loops, special types of arrays and parallelized digital algorithms are supported. The toolbox allows programmer to use the functions supporting parallel calculation with MATLAB and other toolboxes. It is possible to use the toolbox with Simulink® to run several simulations of a model in parallel. The toolbox allows you to use all the processing power of multi-core computers by running applications on workers (MATLAB computing engines) that run locally so each sub-population can be setting up independently on each available worker.

## II. OPTIMIZED PROBLEM

All experiment was realized on Intel Core i5 computers shown as Table I below:

TABLE I. PARAMETERS OF MULTI CORE PROCESSOR

Detail of multi core processor	
Specification	Intel Core i5 8250U CPU @ 1.60 MHz
Core name	Coffee Lake-U/Y
Core Speed	3391.70 MHz
Cores	4
Threads	8



**Rosen Brock : Valley-Shaped**

$$f(x, y) = \sum_{i=1}^n [b(x_{i+1} - x_i)^2 + (a - x_i)^2] \quad (1)$$

Global minimum:  $f(x^*) = f(1, \dots, 1) = 0$

Dimension: 10 ,whereas  $\llbracket X_{1\_limit}, [X_{2\_limit}, \dots, [X_{10\_limit} \rrbracket$

Searchinh region:  $\llbracket -30;30, -30;30, \dots, -30;30 \rrbracket$

**Griewank :Many Local Minima**

$$f(x, y) = 1 + \sum_{i=1}^n \left[ \frac{x_i^2}{4000} - \prod_{i=1}^n \cos \left( \frac{x_i}{\sqrt{i}} \right) \right] \quad (2)$$

Global minimum:  $f(x^*) = f(0, \dots, 0) = 0$

Dim : 20 ,whereas  $\llbracket X_{1\_limit}, [X_{2\_limit}, \dots, [X_{20\_limit} \rrbracket$

Searchinh region:  $\llbracket -600;600, -600;600, \dots, -600;600 \rrbracket$

**Ackley :Many Local Minima**

$$f(x) = f(x_1, \dots, x_n) = -a \cdot \exp \left( -b \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} \right) - \exp \left( -b \sqrt{\frac{1}{n} \sum_{i=1}^n \cos cx_i} \right) + a + \exp(1) \quad (3)$$

Global minimum:  $f(x^*) = f(0, \dots, 0) = 0$

Dim : 20 ,whereas  $\llbracket X_{1\_limit}, [X_{2\_limit}, \dots, [X_{20\_limit} \rrbracket$

Searchinh region:  $\llbracket -30;30, -30;30, \dots, -30;30 \rrbracket$

**Michalewicz :Steep Ridges/Drops**

$$f(x) = f(x_1, \dots, x_n) = -\sum_{i=1}^n \sin x_i \sin^{2m} \left( \frac{ix_i^2}{\pi} \right) \quad (4)$$

Global minimum:  $f(x^*) = -9.66015$

Dimension : 10 ,whereas  $\llbracket X_{1\_limit}, [X_{2\_limit}, \dots, [X_{10\_limit} \rrbracket$

Searchinh region:  $\llbracket 0;\pi, 0;\pi, \dots, 0;\pi \rrbracket$

### III. PROPOSED PARALLEL MULTI-POPULATION TECHNIQUE

**A. Classical PSO**

The standard PSO algorithm initialize a random population of solutions. At each generation, all of fitness function value are calculated to determine best value of those member in this population. In standard PSO, the member of population, called particle, move within a D-dimension searching region with a velocity that is automatically adjusted base on its own experience and sharing information of its neighbors. (Fig. 1)

The particle is represented as  $\vec{x}_i = x_{i1}, x_{i2}, \dots, x_{iD}$ , where  $x_{id} \in [x_{d-\min}, x_{d-\max}]$ ,  $d \in [1, D]$ .  $x_{d-\min}$ ,  $x_{d-\max}$  are the border of searching region of d-th dimension, respectively. The velocity of each particle is represented as  $\vec{v}_i = v_{i1}, v_{i2}, \dots, v_{iD}$  and  $v_{\max}$  is setting up manually. The best

position is calculated at first step of each iterations and stored as  $p_i = p_{i1}, p_{i2}, \dots, p_{iD}$  and compared to best position of all iterations called as global best position represented as  $P_i = P_{i1}, P_{i2}, \dots, P_{iD}$ . At each iteration step, the particles are manipulated according to the following equations:

$$v_i = w \cdot v_i + R_1 \cdot c_1 (P_i - x_i) + R_2 \cdot c_2 (p_i - x_i) \quad (5)$$

$$x_i = x_i + v_i \quad (6)$$

Where  $w$  is inertia weight;  $c_1$  and  $c_2$  are acceleration constant of particle to global and local best position in respectively; and  $R_1$ ,  $R_2$  are random vectors with components uniformly distributed in  $[0,1]$ . In equation (5), the first part of velocity is adjusted by previous experience of itself. The second part is shown that velocity of particle influenced by the global and local best position according to  $c_1$ ,  $c_2$ ,  $R_1$  and  $R_2$  parameter. New velocity vector is updated in each iteration after found out local best position. This process is repeated until reach maximum iteration that is set earlier.

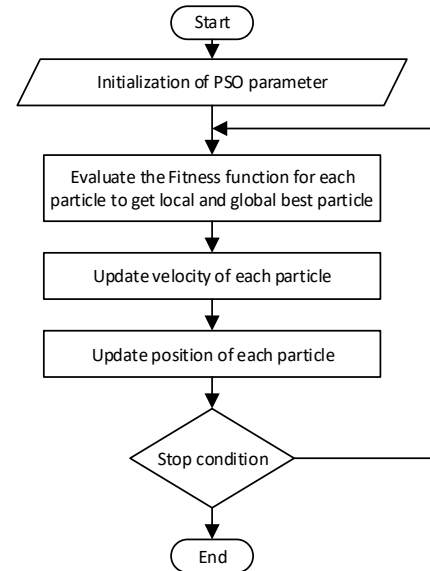


Fig. 1. Flowchart of PSO algorithm

**B. Classical DE**

This section provides an overview of standard Differential Evolution algorithm (DE). This algorithm uses a population of NP candidate solution, each of which is a D-dimensional vector which is confined in specified range (Fig. 2.). The initial population is covered the searching region by randomizing member that its parameter is created in limited range of each dimension. Each member of population is represented as  $\vec{x}_i = x_{i1}, x_{i2}, \dots, x_{iD}$ , where  $i \in [1, NP]$ ,  $x_{id} \in [x_{d-\min}, x_{d-\max}]$ ,  $d \in [1, D]$ .  $x_{d-\min}$ ,  $x_{d-\max}$  are the border of searching region of d-th dimension, respectively. In every iteration, DE perform mutation and crossover operations to produce a mutant vector  $\vec{v}_i = v_{i1}, v_{i2}, \dots, v_{iD}$  and trial vector  $\vec{u}_i = u_{i1}, u_{i2}, \dots, u_{iD}$ .

**1) Mutation.**

A new mutant vector is generated for every individual at every generation using a mutation strategy. The most

frequently used strategies are given in equation (8), where all of members in current generation population  $x_i$  are the shuffled to generate three more population which is its member are represented as  $x_{r1}, x_{r2}, x_{r3}$ . The mutation scale parameter  $F$  should be choose in range  $[0,1]$ .

$$\text{DE/rand/1} \quad v_i = x_{r1} + F * (x_{r2} - x_{r3}) \quad (7)$$

$$\text{DE/best/1} \quad v_i = x_{\text{best}} + F * (x_{r2} - x_{r3}) \quad (8)$$

But DE algorithm used in this paper is chose DE/rand/1 (7) for mutation phase.

### 2) Crossover.

After mutation phase is completed, crossover operation is applied to each target vector  $x_i$  and its corresponding mutant vector  $v_i$  to generate a trial vector  $u_i$ . The binominal crossover used in this phase is given by:

$$u_i = \begin{cases} v_i & \text{rand}(0,1) < CR \\ x_i & \text{otherwise} \end{cases} \quad (1)$$

The crossover factor  $CR$  usually take value within range  $[0,1]$ . This factor is used to control the mutation rate of population.

### 3) Selection.

The selection operation of DE algorithm selects between the current vector  $x_i$  and its corresponding trial vector  $u_i$ . It makes use of a greedy selection.

$$x_i = \begin{cases} u_i & f(u_i) < f(x_i) \\ x_i & \text{otherwise} \end{cases} \quad (2)$$

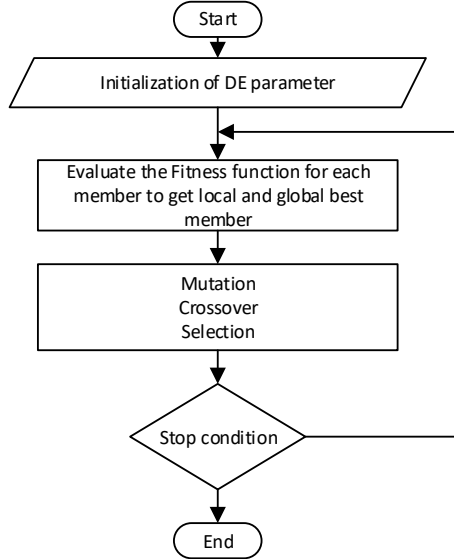


Fig. 2. Flowchart of DE algorithm

### C. Proposed Parallel Multi-Population Technique

In standard PSO or DE algorithm, a population included NP member are created for searching best stage within main searching range  $[[X_{1\_limit}], [X_{2\_limit}], \dots, [X_{n\_limit}]]$ , where  $X_{i\_limit}$  is a limited vector  $[X_{i\_min}, X_{i\_max}]$ ,  $i = 1, n$ . After NG generation, the best stage, where value of member

earned the lowest value of fitness function, is found by NP member.

As Parallel Multi population technique called PM technique, the searching region is split into many sub-searching region according to classified parameter such as  $X_1 \dots X_n$  dimension. Origin searching range is divided into  $k_i$  parts on  $X_i$ -dimension where  $i = 1, \dots, n$ . After all, there are  $k$  sub-searching region where  $k$  is calculated as Equation (11).

$$k = \prod_{i=1}^n k_i \quad (3)$$

Firstly, original population is separated into  $k$  pieces according to  $X_i$ -dimensions shown as Fig. 3. Then, there are two methods to create  $k$  population on those sub-regions. One of two methods is each sub-region is sought for best local stage by NP/k members in NG generations. Another one are each sub-region are searched by NP members within NG/k generation. After all,  $k$  population in  $k$  sub-region return a vector containing best local stage of  $k$  sub-region represented as  $x_{\text{best\_local}} = [x_{\text{best\_1}}, \dots, x_{\text{best\_k}}]$ . Finally, the best global stage is selected by minimum selection strategy as Pseudo code (12).

#### Pseudo code

```

x_best_global = x_best_1;
for i = 1:k
    if f(x_best_i) < f(x_best_global)
        x_best_global = x_best_i;
    end
end
    
```

(4)

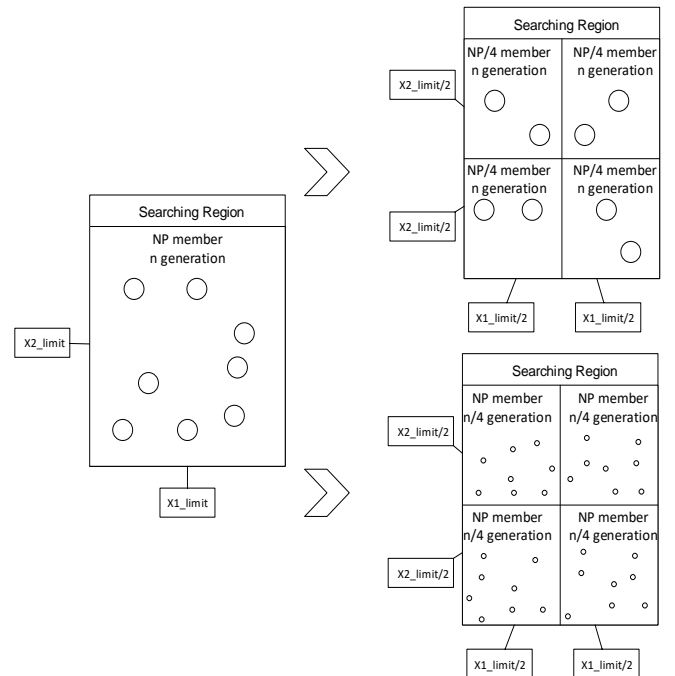


Fig. 3. Principle of Parallel multi population technique

The parallel multi technique expected to perform for fast execution and stable results of optimizing algorithms. Firstly, splitting up searching region decreased local extremes rate of the results because there are  $k$  value. Secondly, the execution of upgraded algorithm takes less

time than the original ones that are shown in experiment results. Finally, the standard deviation and average of results when using parallel multi population technique are competitive compared to standard algorithms.

The performance and effectiveness of parallel multi population are tested on benchmark function detailed in section II. And then, it is compared to the standard algorithms. PSO and DE is two algorithms are tested with parallel multi population. All experiments are simulated by Matlab 2016b on Intel Core i5 8<sup>th</sup> Gen 1.6Mhz and 8Gb RAM.

The control parameters of all algorithms used to optimize are listed in Table II. It is included inertia weight ( $w$ ), Particle's best weight ( $c1$ ), Swarm's best weight ( $c2$ ) for PSO algorithm and mutant factor ( $F$ ), crossover factor ( $CR$ ) for DE algorithm. After few trial simulating, the population and generation of standard PSO and DE algorithm are 50 and 2000 for all benchmark function, respectively. The dimension and the searching range of benchmark function are chosen as equation (1-4) mentioned in section II. The parallel-multi-population using PSO algorithm is used the first method mentioned in Fig. 3 that the searching region is divided into four part according to X1-dimension, the pop size are 50 particals and the generations is 500. The parallel multi population using DE algorithm is used the second method proposed in Fig 3 that the searching region is divided into four part according to X1-dimension, pop size of each sub-searching region is 13 and 2000 generation. Those parameter is chosen to guarantee that the number of computation carried out by processor in each algorithms is equal.

TABLE II. THE PARAMETERS OF PSO, DE, PM-PSO AND PM-DE

	PSO	W=0.8	DE	PM-PSO	W=0.8	PM-DE
		C1=0.4			C1=0.4	
		C2=0.3			C2=0.3	
NP	50		50	50		13
NG	2000		2000	500		2000
K	1		1	4		4
CP	100000		100000	100000		104000

CP: Computatin Cost=NP\*NG\*k

#### IV. RESULTS AND DISSCUSSION

For each above problem, 50 independent runs are carried out and the statistical results of the best, worst, mean and

standard deviation for four optimization algorithms are shown in Table III. And Fig. 5 shows the convergence rate of PM-PSO, PM-DE, PSO and DE algorithms in the optimization of the benchmark functions. The bold values show the best value of each line, which is compared results between classical and parallel multi algorithms, in the result table.

As can be seen from the results in Table III and Fig. 5, the Parallel Multi Technique applied on PSO and DE yields superior results compared to the classical ones in running time aspect. It is clear that the running time of the heuristic algorithms applied PM Technique are significantly faster than the classical ones. In detail, Rosenbrock function optimized by PSO take 0.21s per run meanwhile it take 0.17s in PM technique. Same results as Rosenbrock function, Griewank, Ackley and Michalewicz tested by classical PSO take 0.42s, 0.68s and 0.47s slower than approximately three times when running by PM-PSO that is 0.19s, 0.30s and 0.18s, respectively. When using DE algorithms for those benchmark function, it is obvious that the performing time of DE algorithms is significant faster than PSO algorithms because of number of for loop using in each generation. Above-mentioned principle of both algorithms, PSO have to use two "for loop", one for calculating all member value and one for moving member after that. Whereas, DE algorithms just use 1 "for loop" for selection phase because DE/rand/1 is used in Mutation phase so finding best member is unnecessary. As indicated in Fig. 4., The Parallel Multi Technique is slightly improved processing time of classical DE algorithm and the returned results are completely finer which shown in Table III.

In the case of the optimization of the Griewank, the searching region  $[-600;600]$  is divided into four pieces  $[-600; -300]$ ;  $[-300;0]$ ;  $[0;300]$ ;  $[300;600]$  according to X1-dimension. And the global minimum of those benchmark function is placed on the line that divides origin area so it is hard to find minimum by PSO algorithms. The best value of Griewank found after 50 runs by PSO is 0.0951 lower than running by PM-PSO, 0.2916, three times. PM-PSO also return worse result in Ackley benchmark function compared to traditional ones with 2.7652 and 3.0496, respectively.

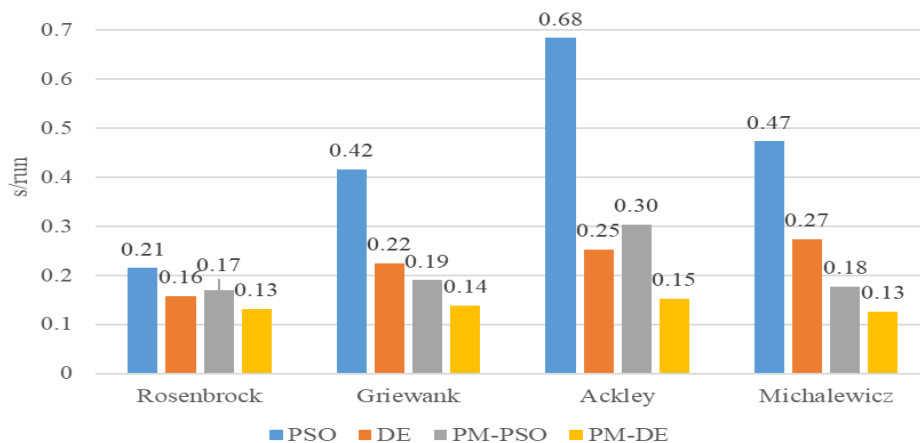


Fig. 4. Running time of PSO/DE compared to PM-PSO/DE

TABLE III. THE EXPERIMENT RESULTS OF TESTING BENCHMARK FUNCTION.

Function		PSO	DE	PM-PSO	PM-DE
Rosenbrock	Best	4.0068	1.8054e-9	<b>0.0017</b>	<b>1.6709e-21</b>
	Worst	3.3315e+4	1.2293e-4	<b>78.9591</b>	<b>1.2915e-5</b>
	Mean	215.1817	2.5610e-6	<b>7.2705</b>	<b>2.6069e-7</b>
	StdDev	556.9688	1.7198e-5	<b>12.0427</b>	<b>1.8078e-6</b>
	Time (s/run)	0.2148	0.1572	<b>0.1697</b>	<b>0.1314</b>
Griewank	Best	0.0951	0.16983	0.2916	<b>7.8826e-15</b>
	Worst	1.2870	0.8442	3.5631	<b>0.0148</b>
	Mean	0.7840	0.6182	1.7050	<b>0.0020</b>
	StdDev	0.3560	0.1752	0.7924	<b>00039</b>
	Time (s/run)	0.4165	0.2245	<b>0.1905</b>	<b>0.1386</b>
Ackley	Best	2.7652	2.0632e-12	3.0496	<b>2.6645e-15</b>
	Worst	10.2113	2.2303e-11	8.8752	<b>2.6645e-15</b>
	Mean	5.7396	7.9714e-12	6.3287	<b>2.6645e-15</b>
	StdDev	1.5309	3.6066e-12	<b>1.3282</b>	<b>0</b>
	Time (s/run)	0.6841	0.2527	<b>0.3033</b>	<b>0.1524</b>
Michalewicz	Best	-8.7674	-9.5301	<b>-9.1561</b>	<b>-9.7562</b>
	Worst	-4.2685	-5.3100	<b>-5.7805</b>	<b>-8.9545</b>
	Mean	-6.3755	-7.6885	<b>-7.3387</b>	<b>-9.3863</b>
	StdDev	1.0549	1.1266	<b>0.7246</b>	<b>0.1692</b>
	Time (s/run)	0.4733	0.2742	<b>0.1767</b>	<b>0.1263</b>

**Bold Value** is the best value in each validate aspect when using PM technique compared to Standard ones.

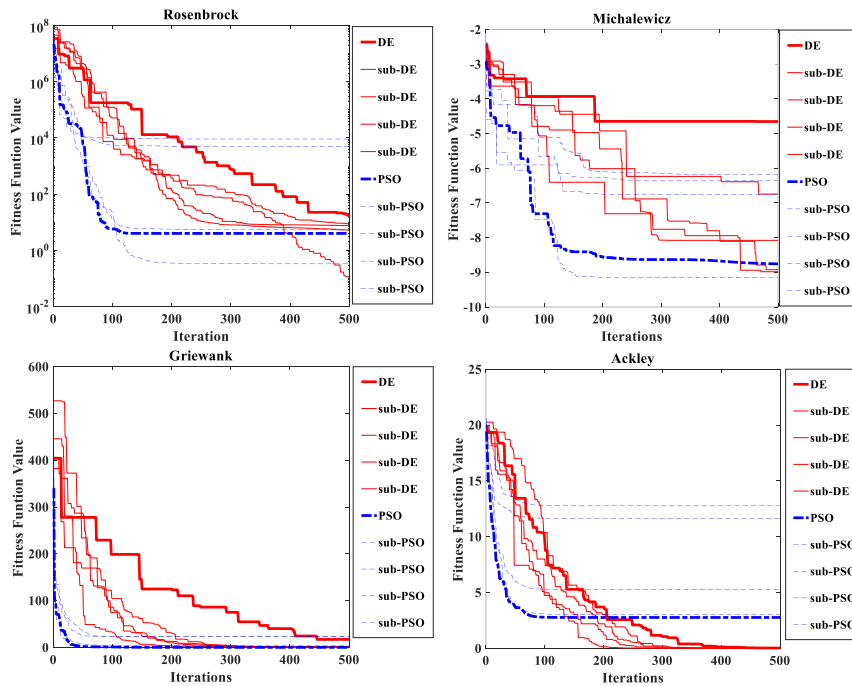


Fig. 5. Convergence rate of PSO,DE,PM-PSO and PM-DE in optimization

From the data in Fig .5, the blue line represent for PSO algorithm and the red line are DE algorithms. The sub-DE and sub-PSO lines are symbolized for sub searching region that split from the origin. As reflected in Fig.5, the convergence rate of PSO is substantially faster than DE. Moreover, almost of sub-DE lines is approached the global minimum more quickly than DE line. Meanwhile, sub-PSOs line is converged at the global minimum at same rate compared to PSO line. Especially, in Rosenbrock and Michalewicz convergence rate figure, the sub-PSO lines,

which contain the global minimum, reach more closer the global minimum than the classical ones.

## V. CONCLUSION

The benchmark function such as Rosenbrock, Griewank, Ackley and Michalewicz is used to test proposed technique. The results show that multi-population technique applied on PSO and DE algorithm has a competitive performance compared to the standard ones. The parallel multi-population technique shown better result which is more

precise and stable. Furthermore, the parallel multi-core (PM) technique applied on PSO and DE optimization process yields superior results compared to the classical ones in running-time requirement. It is clear to note that the running time of the meta-heuristic optimization algorithms applied parallel multi-core (PM) technique shows significantly faster than the classical ones executed on single-core.

#### ACKNOWLEDGMENT

This paper is funded by Vietnam National University of Ho Chi Minh City (VNU-HCM) under grant number B2020-20-04. We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.

#### REFERENCES

- [1] EBERHART, Russell; KENNEDY, James. A new optimizer using particle swarm theory. In: *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. Ieee, 1995. p. 39-43.
- [2] KENNEDY, James; EBERHART, Russell. Particle swarm optimization. In: *Proceedings of ICNN'95-International Conference on Neural Networks*. IEEE, 1995. p. 1942-1948.
- [3] STORN, Rainer; PRICE, Kenneth. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 1997, 11.4: 341-359.
- [4] Deng, W., Yao, R., Zhao, H. *et al.* A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. *Soft Compute* **23**, 2445–2462 (2019).
- [5] Moayedi, H., Mehrabi, M., Mosallanezhad, M. *et al.* Modification of landslide susceptibility mapping using optimized PSO-ANN technique. *Engineering with Computers* 35, 967–984 (2019).
- [6] Moayedi, H., Raftari, M., Sharifi, A. *et al.* Optimization of ANFIS with GA and PSO estimating  $\alpha$  ratio in driven piles. *Engineering with Computers* 36, 227–238 (2020).
- [7] Krell, Evan, *et al.* "Collision-Free Autonomous Robot Navigation in Unknown Environments Utilizing PSO for Path Planning." *Journal of Artificial Intelligence and Soft Computing Research* 9.4 (2019): 267-282.
- [8] Borjigin, Surina, and Prasanna K. Sahoo. "Color image segmentation based on multi-level Tsallis–Havrda–Charvát entropy and 2D histogram using PSO algorithms." *Pattern Recognition* 92 (2019): 107-118.
- [9] Ye, Zipeng, *et al.* "DE-Path: A Differential-Evolution-Based Method for Computing Energy-Minimizing Paths on Surfaces." *Computer-Aided Design* 114 (2019): 73-81.
- [10] Borges, Lucas Camargos, *et al.* "Determination of Nano-aerosol Size Distribution Using Differential Evolution." *Computational Intelligence in Emerging Technologies for Engineering Applications*. Springer, Cham, 2020. 123-136.
- [11] Araújo, Ricardo de A., Germano C. Vasconcelos, and Tiago AE Ferreira. "Improved differential evolution for single objective optimization problems." (2019).
- [12] Liang, Jing, *et al.* "Multimodal multiobjective optimization with differential evolution." *Swarm and evolutionary computation* 44 (2019): 1028-1059.
- [13] TEWOLDE, Girma S.; HANNA, Darrin M.; HASKELL, Richard E. Multi-swarm parallel PSO: Hardware implementation. In: *2009 IEEE Swarm Intelligence Symposium*. IEEE, 2009. p. 60-66.
- [14] KIM, Jong-Yul, *et al.* Optimal power system operation using parallel processing system and PSO algorithm. *International Journal of electrical power & energy systems*, 2011, 33.8: 1457-1461.
- [15] TASOULIS, Dimitris K., *et al.* Parallel differential evolution. In: *Proceedings of the 2004 congress on evolutionary computation (IEEE Cat. No. 04TH8753)*. IEEE, 2004. p. 2023-2029.
- [16] PENAS, David R., *et al.* Enhanced parallel differential evolution algorithm for problems in computational systems biology. *Applied Soft Computing*, 2015, 33: 86-99.
- [17] HUNG, Yukai; WANG, Weichung. Accelerating parallel particle swarm optimization via GPU. *Optimization Methods and Software*, 2012, 27.1: 33-51.



# Adaptive MIMO Fuzzy Controller for Double Coupled Tank System Optimizing by Jaya Algorithm

Cao Van Kien

Faculty of Electronics Technology  
Industrial University of Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
caovankien@iuh.edu.vn

Nguyen Ngoc Son

Faculty of Electronics Technology  
Industrial University of Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
nguyenngocson@iuh.edu.vn

Ho Pham Huy Anh 

Faculty of Electrical & Electronics  
Engineering (FEEE),  
Ho Chi Minh City University of Technology,  
VNU-HCMC  
Ho Chi Minh City, Vietnam  
hphanh@hcmut.edu.vn

**Abstract**— This paper proposes an adaptive MIMO fuzzy controller optimized with Jaya optimization technique used to robust control of uncertain coupled tank system. Firstly, the parameters of MIMO fuzzy controller are optimally identified using Jaya algorithm. Then a fuzzy sliding surface is implemented to ensure that the coupled tank system is asymptotically stable based on Lyapunov concept. The proposed algorithm is applied to control the fluid level of double tank systems. The comparison results with PID and traditional fuzzy controller are presented to confirm that the new Adaptive MIMO Fuzzy control method proves a robust and simple approach to effectively control highly nonlinear uncertain systems.

**Keywords**— Adaptive MIMO Fuzzy controller, coupled tank system, Jaya Optimization Algorithm, Lyapunov Stability Principle, membership function (MF), adaptive fuzzy sliding surface

## I. INTRODUCTION

Fuzzy set was initially introduced in 1965 from Zadeh [1]. Up to now there have been numerous researches improved using this concept, including 2-type Fuzzy, neural fuzzy, hierarchical fuzzy structure, etc. used to identify and regulate uncertain nonlinear plants [2-3]. Nowadays, Takagi-Sugeno (TS) fuzzy set has confirmed its performance in giving an effective scheme for highly nonlinear systems. The advantages of TS fuzzy schemes are that they permit to use a set of local linear fuzzy models with a reduced number of MFs as to successfully represent highly nonlinear systems. Then the TS fuzzy scheme was increasingly used in versatile applications, such as in [4 – 6]. Other developments applying the TS fuzzy model were consulted in [7-9] thereby a TS fuzzy-based algorithm was suggested for uncertain plants with an added  $H_2$ - $H_\infty$  block for well tracking referential trajectories. Nevertheless, in case the MFs of the TS fuzzy structure contain parametric uncertainties, the fuzzy plants cannot handle well. Moreover regarding to a complicated system, it will be needed huge time for estimating with a great of MFs required along with clumsy fuzzy rule-bases.

In order to achieve better accuracy from the fuzzy control approaches, not only its coefficients required to be optimized but the fuzzy scheme is also to be ameliorated. Nowadays, the Type-2 fuzzy structures [10-12] have demonstrated its performance more convincingly than Type-1 in regulating the uncertain plants regarding to noises. Paper [13] presented

an original fuzzy method for uncertain plants using the Type-2 fuzzy model. Paper [14] suggested a new application using Type-2 fuzzy one. Furthermore, many researchers used the meta-heuristic optimization algorithms including a cuckoo search (CSA) [15], Particle Swarm Optimization (PSO) [16], genetic algorithm (GA) [17], differential evolution (DE) [18-19] to optimally identify the coefficients of the fuzzy model in order to efficiently handle the nonlinear features of plants. Apart from the ordinary fuzzy structure, the multi-layer fuzzy set shows that it is very hard to be implemented using only the experience of the designer. So far it is available to be designed by intelligent optimization approaches. Thus the identified multi-layer fuzzy one is ready to be applied to MIMO systems with huge advantage with respect to its well-scaling capability for even complicated plants [20]. Paper [21] successfully introduced a new multi-layer fuzzy structure for modeling a non-linear plant. In [27] a new meta-heuristic called Jaya is suggested. Jaya represents a swarm-based optimization technique whose innovative concept is that an optimum solution of a technical problem usually go away from the bad candidate and, in parallel, searching its path to the best candidate. Furthermore, JAYA gets an advantage over all meta-heuristic optimum approaches since it requires no specific coefficients which are often needed to be sophisticatedly chosen, and hence requires only simple coefficients like swarm size and/or the maximal number of generations.

Nowadays as to remove the uncertain features as to guarantee the stability of highly nonlinear plant, numerous adaptive fuzzy controllers integrated with modern control algorithms for example sliding technique [22-25],  $H_2/H_\infty$  technique [7-9], back-stepping technique [26-27]. Nevertheless, the above-mentioned methods require a sufficient knowledge over the characteristics of investigated uncertain plants. Furthermore, an adaptive fuzzy control method initially uses its parameters in random values that often causes the initial control procedure strongly hard to be handled, then it eventually issues with overshoot and long transient-time response.

To surpass such above-mentioned disadvantages, this study initiatively introduces a new adaptive multilayer fuzzy control (AMFC) approach for controlling an uncertain nonlinear MIMO plants.

## II. SYSTEM MODELLING

Investigated plant is a typical uncertain MIMO one which have 2 inputs (Pump 1 and 2), 2 fluid-level outputs ( $x_2, x_4$ ) with its structure is describes in Fig. 1:

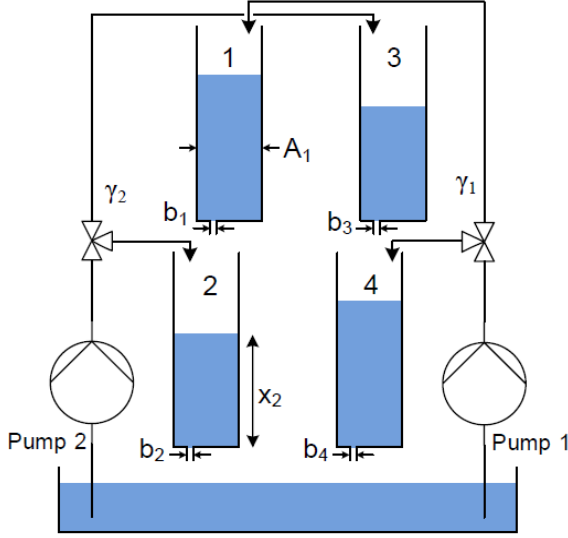


Fig. 1. Investigated Double- Coupled Tank System

Figure 2 shows that pump-1 directly controls tank-1 fluid level, tank-2 was affected by tank-1 outlet, pump 2 directly controls tank-3 fluid level, and tank-4 was affected by tank-3 outlet. It means that an interaction exists among  $u_1$  and  $x_4$ ,  $u_2$  and  $x_3$ . Consequently that coupled effect makes system strongly difficult to precisely control the fluid-level.

The dynamic functions of the investigated plant can be described via weight-balancing function and Bernoulli's concept, as follows:

$$\begin{cases} \frac{dx_1}{dt} = \frac{K\gamma_1 u_1(t)}{A_1} - \frac{b_1 C \sqrt{2gx_1}}{A_1} \\ \frac{dx_2}{dt} = \frac{K(1-\gamma_2)u_2(t)}{A_2} + \frac{b_1 C \sqrt{2gx_1}}{A_2} - \frac{b_2 C \sqrt{2gx_2}}{A_2} \\ \frac{dx_3}{dt} = \frac{K\gamma_2 u_2(t)}{A_3} - \frac{b_3 C \sqrt{2gx_3}}{A_3} \\ \frac{dx_4}{dt} = \frac{K(1-\gamma_1)u_1(t)}{A_4} + \frac{b_3 C \sqrt{2gx_3}}{A_4} - \frac{b_4 C \sqrt{2gx_4}}{A_4} \end{cases} \quad (1)$$

with  $u_1, u_2$  represents voltage control of pump motor 1, 2.  $x_1, x_2, x_3, x_4$  denote fluid levels of Tank 1-4. The full definitions and configurations are presented in Table 1. The selected parameters are close to the real model parameters

TABLE 1. PHYSICAL MEANING AND NUMERICAL VALUE USED IN THE EXPERIMENT

Notation	Physical meaning	Value[unit]
$A_1$	Tank 1 interior diameter	16.619(cm <sup>2</sup> )
$A_2$	Tank 2 interior diameter	16.619(cm <sup>2</sup> )
$A_3$	Tank 3 interior diameter	16.619(cm <sup>2</sup> )
$A_4$	Tank 4 interior diameter	16.619(cm <sup>2</sup> )

Notation	Physical meaning	Value[unit]
$b_1$	Outflow orifice diameter of Tank 1	0.5 (cm <sup>2</sup> )
$b_2$	Outflow orifice diameter of Tank 2	0.4(cm <sup>2</sup> )
$b_3$	Outflow orifice diameter of Tank 3	0.5(cm <sup>2</sup> )
$b_4$	Outflow orifice diameter of Tank 4	0.4(cm <sup>2</sup> )
$C$	The discharge outlet parameter	0.8
$g$	Gravitation	981(cm/s <sup>2</sup> )
$K$	Pump flow value	6.94(cm <sup>3</sup> /(s.V))
$\gamma_1$	Flow-ratio in tank 1 to tank 4	90(%)
$\gamma_2$	Flow-ratio in tank 2 to tank 3	90(%)

Assuming the system (1) can be rewritten as:

$$\begin{cases} \dot{x}_1 = f_1(x, t) + g_1(x, t)u_1 \\ \dot{x}_4 = f_2(x, t) + g_2(x, t)u_2 \end{cases}$$

With  $f_1, f_2, g_1, g_2$  are assumed uncertainty nonlinear functions including noises and disturbances, the function  $f(x, t)$ ,  $g(x, t)$  were bounded and  $0 < \underline{g} \leq g(x, t) \leq \bar{g} < +\infty$ ,  $|\dot{g}(x, t)| \leq G_g$  for all  $x \in R^n$ .

The main purpose of this study is to design an adaptive fuzzy law for  $u_1, u_2$  outputs as to precisely tracking a reference trajectory regarding to bounded derivatives  $y_{d1}$  and  $y_{d2}$ .

## III. PROPOSED CONTROLLER

This subsection proposed the new AMFC algorithm illustrated in Figure 2 which combines the Multi-layer Fuzzy structure, the adaptive T-S Fuzzy model and the stabilizing control law. The proposed method has advantages over classical adaptive control algorithms because it successfully links both optimal algorithm and Lyapunov stability principle used in control theory. The optimal training helps to reduce the computational cost of the adaptive stage, the additional adaptive algorithm improves the quality of the control algorithm during parameters varied in operation and lastly a perfect inverse controller model is created although it is really hard to do it.

The control law presented in Eq. (2) consists of 3 main elements in which the 1st of the inverse fuzzy rule ( $u_{ifm}^*$ ), the 2nd of the adaptive fuzzy component ( $\theta_u^T \xi(x)$ ) and the last related to the stable control value ( $u_{sw}$ ).

$$u = u_{ifm}^* + u_{af} + u_{sw} \quad (2)$$

with  $u_{ifm}^*$  denotes the optimum inverse AMFC output which is optimally identified with DE and described as

$$u_{ifm}^* = \arg \min_{u_{ifm} \in \Omega_u} \sup |u_{ref} - u_{ifm}| \quad (3)$$

where  $\Omega_u$  represents constraint sets for  $u_{ifm}$ ;

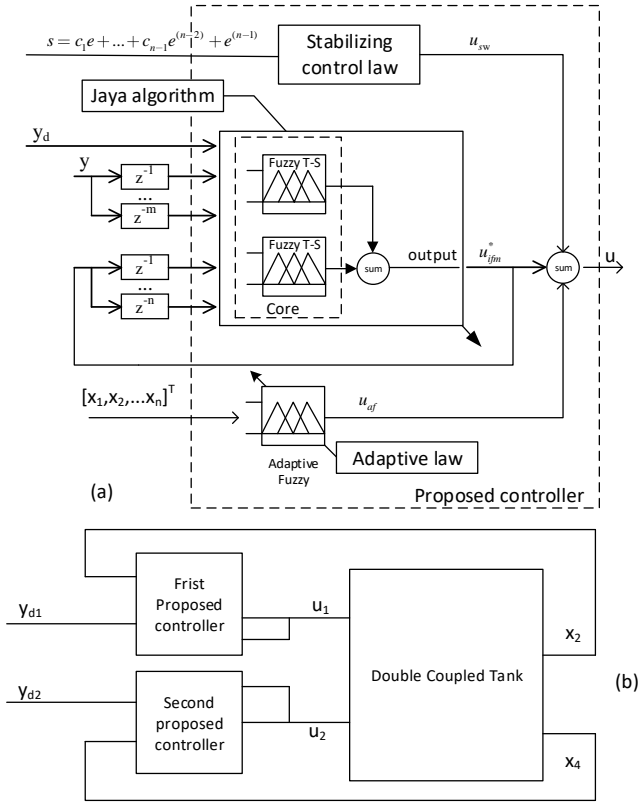


Fig. 2. Block scheme of novel inverse fuzzy AMFC control

$u_{af}$  denotes the output of proposed controller that can be described as,

$$u_{af}(x) = \frac{\sum_{n=1}^N \mu_n(x) \theta_n}{\sum_{n=1}^N \mu_n(x)} \quad (4)$$

where  $\mu_n(x) = \prod_{i=1}^n \mu_{F_i^{ji}}(x_i)$ .

In which  $\mu_{F_i^{ji}}(x_i)$  denotes the membership function (MF) of the fuzzy set  $F_i^{ji}$ , and  $\theta_n$  denotes the point at that the MF attains its maximal value.

Then the output  $u$  is adjusted in followed compact form  $u_{af}(x) = \theta^T \xi(x)$  with  $\theta = \theta_1, \dots, \theta_N^T$  represents vector containing consequent coefficients,  $\xi(x) = \xi_1(x), \dots, \xi_N(x)^T$  represents fuzzy-consequent relations which is calculated as  $\xi_k(x) = \frac{\mu_k(x)}{\sum_{n=1}^N \mu_n(x)}$ ,  $k=1, \dots, N$ . Thus the plant is

assumed to be fitted in term that  $\sum_{n=1}^N \mu_n(x) \neq 0$  with  $x \in U$ . The fuzzy rule is then followed as

$$\dot{\theta}_u = \alpha^{-1} s \xi(x) \quad (5)$$

where  $s$  represents sliding surface;  $u_{sw}$  is designed as

$$u_{sw} = K \text{sign}(s) \quad (6)$$

The fact is that the  $\text{sign}()$  equation is to be changed by  $\text{sat}()$  one as to attenuate the chattering effect.

The optimal identified model-based control had to ensure the closed-loop plant stability. But, it might not be feasible to attain accurate tracking. As to solving it, the purpose aims to obtain the absolute bounded error inside a preset value, that is  $\|e(t)\| < \varepsilon$ , in which  $\varepsilon$  denotes a preset petit positive magnitude.

The adaptive law and  $u_{sw}$  are implemented as to guarantee that the investigated plant ensures an asymptotical stability.

**Proof:**

First, the authors define  $u^*$  which represents ideal control law.

$$u^* = u_{ifm}^* + u_{af}^* \quad (7)$$

where  $u_{ifm}^*$  represents ideal optimal inverse multilayer control signal,  $u_{af}^* = \theta_u^{*T} \xi(x)$  is the ideal adaptive control law. From (2) to (7), it gives:

$$u - u^* = \tilde{\theta}_u^T \xi(x) + u_{sw} \quad (8)$$

with  $\tilde{\theta}_u = \theta_u - \theta_u^*$  represents the error of the adaptive coefficient of the fuzzy rule.

The purpose of this paper is to design a robust fuzzy law as to ensure  $y$  well following the reference trajectory  $y_d$ . Then the error  $e$  is defined in (9):

$$e = y_d - y = [e, \dot{e}, \dots, e^{(n-1)}] \in R^n \quad (9)$$

The sliding surface is calculated as:

$$s = c_1 e + c_2 \dot{e} + \dots + c_{n-1} e^{(n-2)} + e^{(n-1)} \quad (10)$$

Whereas  $c = c_1, c_2, \dots, c_{n-1}, 1$  represent the parameters satisfying the Routh–Hurwitz stability criterion. The  $\dot{s}$  signal denotes the derivation of  $s$  and is determined as:

$$\dot{s} = c_1 \dot{e} + c_2 \ddot{e} + \dots + c_{n-1} e^{(n-1)} + e^{(n)} \quad (11)$$

$$\dot{s} = \sum_{i=1}^{n-1} c_i e^{(i)} + y_d^{(n)} - y^{(n)} \quad (12)$$

$$\dot{s} = f(x, t) + g(x, t) u^* - (f(x, t) + g(x, t) u) + \sum_{i=1}^{n-1} c_i e^{(i)} \quad (13)$$

$$\dot{s} = -g(x) \tilde{\theta}_u^T \xi(x) - g(x) u_{sw} + \sum_{i=1}^{n-1} c_i e^{(i)} \quad (14)$$

Now the Lyapunov function candidate is chosen as in (15),

$$V = \frac{1}{2g(x)}s^2 + \frac{1}{2}\tilde{\theta}_u^T \alpha \tilde{\theta}_u \quad (15)$$

$$\begin{aligned} \dot{V} &= \frac{s\dot{s}}{g(x)} - \frac{\dot{g}(x)s^2}{2g^2(x)} + \tilde{\theta}_u^T \alpha \dot{\theta}_u \\ &= s \left[ -\tilde{\theta}_u^T \xi(x) - u_{sw} + \frac{\sum_{i=1}^{n-1} c_i e^{(i)}}{g(x)} - \frac{\dot{g}(x)s}{2g^2(x)} \right] + \tilde{\theta}_u^T \alpha \dot{\theta}_u \quad (16) \\ &= \tilde{\theta}_u^T \left[ \alpha \dot{\theta}_u - s \xi(x) \right] - s \left[ u_{sw} - \frac{\sum_{i=1}^{n-1} c_i e^{(i)}}{g(x)} + \frac{\dot{g}(x)s}{2g^2(x)} \right] \end{aligned}$$

Using (16), fuzzy rule is adaptively chosen based on Eq.(5-6),

$$\dot{\theta}_u = \alpha^{-1} s \xi(x)$$

$$u_{sw} = K \text{sign}(s)$$

with  $K$  represents stability coefficient,  $\alpha$  is learning parameter, which represent positive variable and is

optimally chosen.  $K$  is chosen as  $K > \left| \frac{\sum_{i=1}^{n-1} c_i e^{(i)}}{g(x)} - \frac{\dot{g}(x)s}{2g^2(x)} \right|$

or for simplicity, the authors choose

$$K > \left| \frac{\sum_{i=1}^{n-1} c_i e^{(i)}}{g(x)} \right| + \left| \frac{\dot{g}(x)s}{2g^2(x)} \right| \text{ or } K > \frac{|\dot{s} - e^{(n)}|}{g} + \frac{G_g |s|}{2g^2}.$$

Then  $\dot{V} \leq 0$ . Using Lyapunov stability concept,  $s, \tilde{\theta}_u$

shows asymptotically stable and the nonlinear system (1) is stable. Selecting  $K$  shown, the authors note that if the inverse model is modeled more precise, the  $K$  value is able to be selected smaller.

#### IV. SIMULATION RESULTS

In this paper, the fuzzy controller for each pump motor has 2 inputs with 5 MFs in Gaussian, 7 output firing strength and total 25 rules. The 7 firing strength are optimized by Jaya algorithm with the cost function selected as follows:

$$J = \sum_{i=1}^N e_{1i}^2 + e_{2i}^2 \quad (17)$$

Where  $e_1, e_2$  are the error between the reference signal and the output of the coupled tanks. The algorithm will find fuzzy model parameters such that the cost function is minimal. Then the adaptive algorithm is applied as in Section 3.

The proposed algorithm is compared with the optimal fuzzy algorithm trained by the Jaya algorithm. The results shown in Fig. 4 and 5 show that the proposed algorithm gives a much better control performance than the optimal control algorithm at the start-up stage.

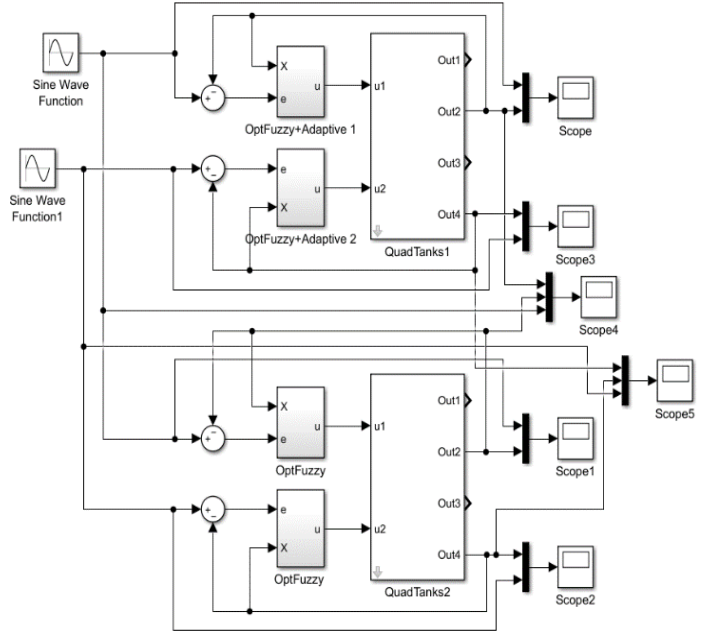


Fig. 3. Control scheme in Matlab/Simulink environment

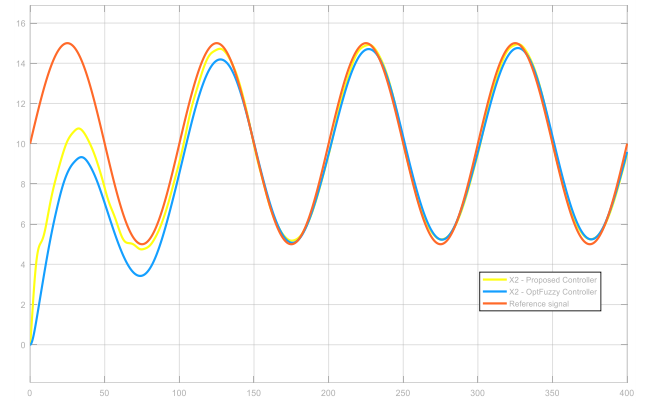


Fig. 4. Response result of the output  $x_2$  of the double coupled tank model

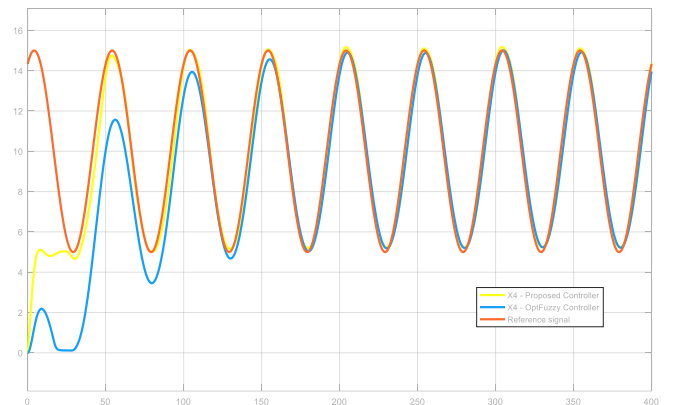


Fig. 5. Response result of the output  $x_4$  of the double coupled tank model

#### V. CONCLUSIONS

The authors in this paper introduce a novel multi-layer fuzzy (AMFC) controller, optimized by Jaya technique, used

for control of uncertain nonlinear MIMO plants. The proposed AMFC is structured based on the optimal multilayer fuzzy model optimally identified by Jaya algorithm with an additional fuzzy sliding surface applied to ensure the asymptotical stability of investigated system using Lyapunov stability concept.

The benchmark test results guarantee that proposed AMFC approach is efficiently applied in control of uncertain MIMO systems. This AMFC method demonstrates quite robust than other traditional fuzzy control approach. As a consequent a scalable multilayer fuzzy model is implemented to successfully control fluid-level of an uncertain coupled-tank system. The limitation of the method is to find an optimal model that can control the system whose identification process can take a long time with respect to complex models. However, for a complex system, optimum identified model integrated in AMFC controller suggested in this paper is an advantage over other traditional adaptive control methods often starting with random parameters. Thus, these results convincingly confirm that AMFC approach should efficiently control of other highly uncertain MIMO plants in further studies.

#### ACKNOWLEDGMENT

This paper is funded by Vietnam National University of Ho Chi Minh City (VNU-HCM) under grant number B2020-20-04. We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.

#### REFERENCES

- [1] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 8, pp. 338-3365, 1965.
- [2] Precup, Radu-Emil, and Hans Hellendoorn. "A survey on industrial applications of fuzzy control." *Computers in Industry* 62.3 (2011): 213-226.
- [3] Qiu, Jianbin, Huijun Gao, and Steven X. Ding. "Recent advances on fuzzy-model-based nonlinear networked control systems: A survey." *IEEE Transactions on Industrial Electronics* 63.2 (2016): 1207-1217.
- [4] Chiu, Chian-Song. "TS fuzzy maximum power point tracking control of solar power generation systems." *IEEE Transactions on Energy Conversion* 25.4 (2010): 1123-1132.
- [5] Rezaee, Babak, and MH Fazel Zarandi. "Data-driven fuzzy modeling for Takagi-Sugeno-Kang fuzzy system." *Information Sciences* 180.2 (2010): 241-255.
- [6] Cheung, Ngaam J., Xue-Ming Ding, and Hong-Bin Shen. "OptiFel: A convergent heterogeneous particle swarm optimization algorithm for Takagi-Sugeno fuzzy modeling." *IEEE Transactions on Fuzzy Systems* 22.4 (2014): 919-933.
- [7] Tseng, Chung-Shi, Bor-Sen Chen, and Huey-Jian Uang. "Fuzzy tracking control design for nonlinear dynamic systems via TS fuzzy model." *IEEE Transactions on fuzzy systems* 9.3 (2001): 381-392.
- [8] Xiaodong, Liu, and Zhang Qingling. "New approaches to  $H_\infty$  controller designs based on fuzzy observers for TS fuzzy systems via LMI." *Automatica* 39.9 (2003): 1571-1582.
- [9] Nasiri, Alireza, et al. "Reducing conservatism in  $H_\infty$  Robust State Feedback Control design of TS Fuzzy Systems: A Non-monotonic Approach." *IEEE Transactions on Fuzzy Systems* (2017).
- [10] Mendel, Jerry M., and RI Bob John. "Type-2 fuzzy sets made simple." *IEEE Transactions on fuzzy systems* 10.2 (2002): 117-127.
- [11] Mendel, Jerry M. "General type-2 fuzzy logic systems made simple: a tutorial." *IEEE Transactions on Fuzzy Systems* 22.5 (2014): 1162-1182.
- [12] Li, Hongyi, et al. "Control of nonlinear networked systems with packet dropouts: interval type-2 fuzzy model-based approach." *IEEE Transactions on Cybernetics* 45.11 (2015): 2378-2389.
- [13] Kumbasar, Tufan, et al. "Type-2 fuzzy model based controller design for neutralization processes." *ISA transactions* 51.2 (2012): 277-287.
- [14] Kumbasar, Tufan, et al. "An inverse controller design method for interval type-2 fuzzy models." *Soft Computing* 21.10 (2017): 2665-2686.
- [15] Berrazouane, Sofiane, and Kamal Mohammadi. "Parameter optimization via cuckoo optimization algorithm of fuzzy controller for energy management of a hybrid power system." *Energy conversion and management* 78 (2014): 652-660.
- [16] Soufi, Youcef, Mohcene Bechouat, and Sami Kahla. "Fuzzy-PSO controller design for maximum power point tracking in photovoltaic system." *International Journal of Hydrogen Energy* 42.13 (2017): 8680-8688.
- [17] Sundarabalan, C. K., and K. Selvi. "Real coded GA optimized fuzzy logic controlled PEMFC based Dynamic Voltage Restorer for reparation of voltage disturbances in distribution system." *International Journal of Hydrogen Energy* 42.1 (2017): 603-613.
- [18] Lutz, Adam, Vlad Bonderev, and Chester Fiesnski. "Fuzzy neural network optimization and network traffic forecasting based on improved evolution." *Neural Networks & Machine Learning* 1.1 (2017): 2-2.
- [19] Chen, Cheng-Hung, and Chong-Bin Liu. "Reinforcement Learning-Based Differential Evolution With Cooperative Coevolution for a Compensatory Neuro-Fuzzy Controller." *IEEE Transactions on Neural Networks and Learning Systems* (2017).
- [20] Tu, Kuo-Yang, Tsu-Tian Lee, and Wen-Jieh Wang. "Design of a multilayer fuzzy logic controller for multi-input multi-output systems." *Fuzzy sets and systems* 111.2 (2000): 199-214.
- [21] C. Van Kien, N. N. Son, and H. P. H. Anh, "Identification of 2-DOF Pneumatic Artificial Muscle System with Multilayer Fuzzy Logic and Differential Evolution Algorithm," in *The 12th IEEE Conference on Industrial Electronics and Applications (ICIEA 2017)*, 2017, pp. 1261-1266.
- [22] Xue, Yanmei, Bo-Chao Zheng, and Xinghuo Yu. "Robust sliding mode control for TS fuzzy systems via quantized state feedback." *IEEE Transactions on Fuzzy Systems* (2017).
- [23] Li, Hongyi, et al. "Adaptive sliding mode control for Takagi-Sugeno fuzzy systems and its applications." *IEEE Transactions on Fuzzy Systems* (2017).
- [24] Li, Jinghao, et al. "Observer-Based Fuzzy Integral Sliding Mode Control For Nonlinear Descriptor Systems." *IEEE Transactions on Fuzzy Systems* (2018).
- [25] Tong, Shaocheng, et al. "Observer-based adaptive fuzzy backstepping control for a class of stochastic nonlinear strict-feedback systems." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.6 (2011): 1693-1704.
- [26] Li, Hongyi, et al. "Adaptive fuzzy back-stepping tracking control for strict-feedback systems with input delay." *IEEE Transactions on Fuzzy Systems* 25.3 (2017): 642-652.
- [27] Rao RV. Jaya : A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *Int J Ind Eng Comput*, Vol. 7, pp. 19-34, 2016.



# A New Approach for Analyzing and Predicting Carbon Dioxide Emissions: Case study of Vietnam

Le Thi Giang  
Faculty of Commerce  
University of Finance -  
Marketing  
HCM city, Vietnam  
lenthigiang@ufm.edu.vn

Khuu Manh Dat  
Faculty of Commerce  
University of Finance -  
Marketing  
HCM city, Vietnam  
minhdat@ufm.edu.vn

Nguyen Xuan Hiep  
Faculty of Commerce  
University of Finance -  
Marketing  
HCM city, Vietnam  
nxhiep@ufm.edu.vn

Sam Nguyen-Xuan  
Faculty of Information  
Technology  
Posts and Telecoms Institute  
of Technology  
HCM city, Vietnam  
samnx@ptithcm.edu.vn

**Abstract**— Vietnam has the potential to affect to air pollution such as increasing carbon-dioxide emission from the rapid extension of gross domestic products, foreign direct investment, and manufacturing sections. Thus, this work analyzes the relationships between the variables and proposed a dynamic model for predicting carbon dioxide emissions in Vietnam based on the relationships. Moreover, the work presents trend prediction of carbon dioxide emissions in Vietnam based on multiple regression model. The work shown that there is a statistically significant positive association between carbon-dioxide emission and manufacturing. However, the other economic sectors, gross domestic product, and foreign direct investment, are weaker impact on carbon-dioxide emission.

**Keywords**—multiple regression (MR); gross domestic products (GDP); manufacturing (MAN); foreign direct investment (FDI); carbon-dioxide emission (CO<sub>2</sub>E).

## I. INTRODUCTION

Recently, many developing countries all over the world are looking for a new economic model which addresses the issue of economic growth and environmental protection. Therefore, they need an approach, which figure out the relationship between economic growth and environmental protection. Basically, the study is known as the green growth model (GGM) and it is adopted by the ministerial conference on environment and development (MCED) [1]. The approach is complex because several economic sectors come into play, including carbon-dioxide emission, manufacturing, gross domestic product, foreign direct investment, and so on.

The GGM model plays an important role to achieve equal pay because it allows Vietnamese government pay attention on investments in industrial development as well as environmental protection. To improve economic growth rapidly, the present suggestions focus on maximizing the benefit of industrial sector while light on the other sectors. The policies have led to the rapidly increasing energy consumption such as fossil fuel, natural gas, etc. In [2], carbon-dioxide emission (CO<sub>2</sub>E) is produced significantly due to the manufacturing activity. Caused by environmental degradation problems, the quality of Vietnam's growth has declined. Therefore, Vietnam must take greater efforts to *improve policies and accurate measures* the relationship between CO<sub>2</sub>E and economic growth.

However, the trade conflict between China and the US has the potential to affect to extension of industry section of Vietnam's economy because foreign manufacturers in China have been moving in Vietnam. While this trend can become a good chance for economic growth of Vietnam, it has the potential to affect to air pollution such as increasing carbon-dioxide emission from the rapid extension of industry section. Examining impacts of the trend in advance is necessary to adapt current economic model relating with environmental sustainability in Vietnam.

In general, the relationship between CO<sub>2</sub>E and economic growth can be modelled by the environmental Kuznets curve [3, 4]. The environment Kuznets curve (EKC) shows an inverted U shape relationship between economic growth and environmental degradation. In [5], a study examined relationship among CO<sub>2</sub>E, income, energy consumption, trade openness, financial development, and so on for 151 countries. The study examination is followed by ECK model [4].

To clear understanding the relationship between CO<sub>2</sub>E and gross domestic products (GDP), manufacturing (MAN) and foreign direct investment (FDI) sectors, an analysis of the relationship must show. The aim of the work is to examine is there any a significant and positive correlation between CO<sub>2</sub>E and the other sectors and a significant increase in CO<sub>2</sub>E due to FDI and MAN. Because data for this work is collected from (The World Bank data), however, the sample periods of the resource for CO<sub>2</sub>E, GDP, MAN, FDI are different. Thus, we proposed new approach to optimal the data by varying  $K$  values.

In addition, we analyze and investigate correlation coefficients for each  $K$  value. Each scenario is used to reveal relationship between CO<sub>2</sub>E and other sectors. Moreover, predicting of CO<sub>2</sub>E in the future for each  $K$  value will be shown and compared with business as usual (BAU) in the work. The contribution make our proposed model is different to previous works [5-9]. Take the results into account, Vietnamese government may pay more attention on investments in industrial development as well as environment problems.

The rest of this paper is organized as follows: Section 2 introduces the related works, section 3 is the methodology framework of our proposal, then section 4 is our results on different scenarios, and the last section shows conclusions and some policy implications are provided.

## II. RELATED WORKS

### A. Economy affects CO<sub>2</sub> emissions

In [10], economic activity has divided into three sectors which are agriculture, industry, and service. The agriculture sector of the economy is the sector of an economy making direct use of natural resources. The sector is the process of producing food, feed, fiber, and other goods by the systematic raising of plants and animals. In contrast, industry is the sector of economy concerned with production of goods including fuels and fertilizers. The last one, the service sector, produces services. The agriculture sector is usually most important in less-developed countries and typically less important in industrial countries and all of them are a component of the GDP of a nation.

The economic growth leads to a shift from agriculture sector to industry sector, which may consume natural resources and contribute CO<sub>2</sub>E. Since the beginning of the industrial revolution have produced large percentage increase in the atmospheric concentration of carbon dioxide [11, 12]. This implies that rapid economic growth is necessary to tackle poverty, and problems related with basic standard of life. However, the shifting economic from agriculture sector to industrial sector caused the depletion of natural resource and the deterioration of the environment.

A consequence of the rapid economic growths generated unsustainable development. This has threatened our life and survival of generations in the future. Until now, there had been struggling with definition of sustainability. In [13], sustainable development is defined as development that meets "the needs of the present without compromising the ability of future generations to meet their own needs". Therefore, to achieve both economic growth and environmental protection, international community has sought effective ways to make sustainable growth.

### B. Context of Vietnam

After renovation and opening market, the average CO<sub>2</sub>E of Vietnam increased more than five times from 1986 to 2014. As the most important engines after renovating and opening market, FDI and manufacturing hold the main role for economic growth in Vietnam, however, the engines are not only effect on GDP but also contribute to the rising of CO<sub>2</sub>E. Given growing concerns over CO<sub>2</sub>E, and the climate change effects, analyzing and predicting of the engines contribute to CO<sub>2</sub>E will provide recommendations and suggestions for long-term policies to implement frameworks for foreign trade actions. Therefore, the answers to questions such as how strong the engines contribute to the rising of CO<sub>2</sub>E and what is the relationship among them are critical metrics for resolving above concerns.

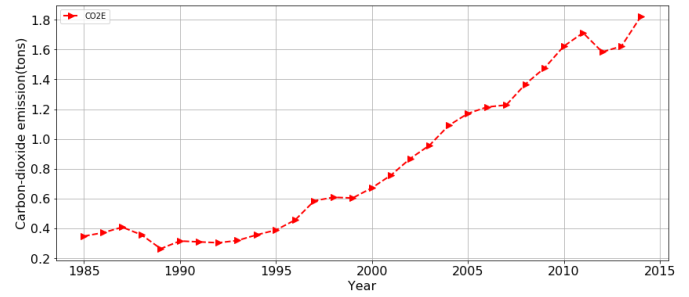


Fig. 1. Carbon dioxide emissions of Vietnam from 1986 to 2014 [14].

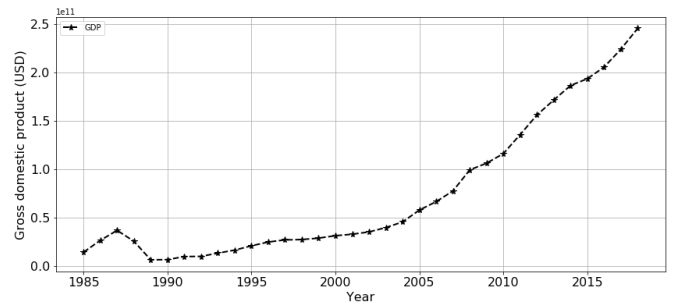


Fig. 2. The economic growth in Vietnam from 1986 to 2018 [14].

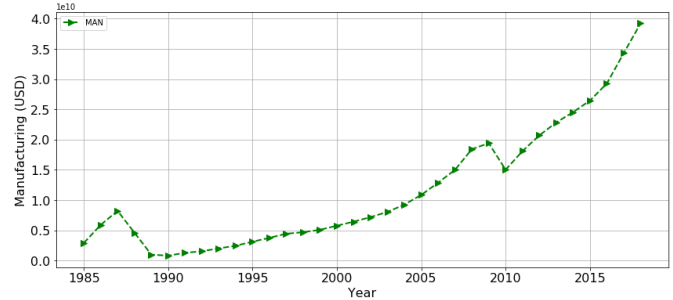


Fig. 3. Manufacturing of Vietnam from 1986 to 2018 [14].

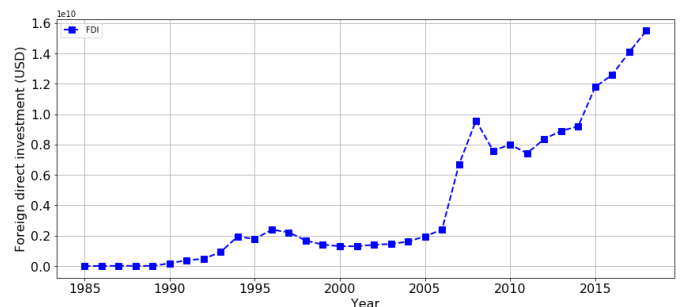


Fig. 4. The FDI inflows of Vietnam from 1986 to 2018 [14].

We consider historical values of all sections. As shown in Fig. 1 and Fig. 2, the CO<sub>2</sub>E and economic growth have been rising rapidly within 30 years. Thus, the Vietnam government has been focusing on the problems and aims to reduce GHG emissions by 8 percent below the business as usual (BAU) scenario by 2030 [4]. On the other hand, the effect of FDI and manufacturing also increases gradually with the increase of the CO<sub>2</sub>E, these are showed in Fig. 3 and Fig. 4. Over long-term outlook to achieve BAU for CO<sub>2</sub>E, there is a relationship among CO<sub>2</sub>E, FDI, and manufacturing in which the coefficients of FDI and manufacturing will indicate how strong the engines impact on CO<sub>2</sub>E.

## III. METHODOLOGY

## A. Multiple regression equations

In [15], there are several models which presented the relationship between economic growth, energy, and CO<sub>2</sub> emissions. To express a relation between CO<sub>2</sub>E and GDP, manufacturing, and FDI, a mapping from the sectors to CO<sub>2</sub>E in which CO<sub>2</sub>E is a dependent variable and GDP, manufacturing, and FDI are independent variables. We propose a modeled form, shown in equation (1) as following:

$$CO_2E = f(GDP, MAN, FDI) \quad (1)$$

where, CO<sub>2</sub>E is denoted as values of CO<sub>2</sub> emission (tons), GDP observation represents for economic growth, MAN is manufacturing refers to industries, and FDI refers to direct investment equity flows. Given any value of the predictors, we can estimate the CO<sub>2</sub>E by taking the logarithmic form on both sides using the multiple regression model [16]. The equation (1) is then rewritten into logarithm as following:

$$\ln(\hat{y}) = \beta_0 + \beta_1 \ln(GDP) + \beta_2 \ln(MAN) + \beta_3 \ln(FDI) \quad (2)$$

where  $\beta_0$  is regression constant, and  $\beta_i$  ( $i = 1, 2, 3$ ) is the regression coefficient to be determined from the variable factors of GDP, MAN, and FDI.

There are several ways to find the coefficients for the best multiple regression, the most common method to measure closeness is to minimize the residual sum of squares (*rss*). Thus, the difference between the  $i^{th}$  expected value  $\hat{y}_i$  and the  $i^{th}$  predicted value  $y_i$  is presented  $i^{th}$  residual,  $\epsilon_i = y_i - \hat{y}_i$ . We define the *rss* as:

$$rss(\beta_i) = \min(\sum_{i=1}^n \epsilon_i^2) \quad (3)$$

where  $\epsilon_i$  ( $i = 1, 2, 3$ ) a vector of residual terms.

As showed in equation (3), if *rss* is small, it means that the model predictions are very close to the actual values and versa. To solve the optimization problem, we transformed the problem into linear equations by using matrix operations. Generally, the equation (3) is equivalent as:

$$rss(\beta_i) = \sum_{i=1}^n (y - X\beta)^T (y - X\beta) \quad (4)$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix}$$

where X is data matrix with an extra column of ones on the left to account for the intercept,  $y = (y_1, \dots, y_n)^T$ , and  $\beta = (\beta_1, \dots, \beta_n)^T$ .

## B. Data Set and Sample Period

The data for this work is collected from world bank data [14] where the sample period is taken into consideration for GDP, manufacturing, and FDI from 1986 to 2018 and the sample period is taken into consideration for CO<sub>2</sub>E from 1986 to 2014. Table I shows the summary of the variables used in the multiple regression equations. In the table I, CO<sub>2</sub>E

has a unit measurement of tons per capita, GDP is gross domestic product per capita in US dollars, MAN is manufacturing added value in US dollars, and FDI represents per capita of foreign direct investment in US dollars.

TABLE I. THE SUMMARY OF THE VARIABLES FROM 1986 TO 2018.

Year	CO <sub>2</sub> E	GDP	MAN	FDI
1986	0.3706802	26336616250	5891663779	40000
1987	0.407674883	36658108850	8200176618	10363703.7
1988	0.355998429	25423812649	4590135743	7680000
1989	0.263108391	6293304975	953639655.7	4070000
1990	0.31487431	6471740806	793175839.4	180000000
1991	0.308941403	9613369520	1259650117	375190278
1992	0.302998931	9866990236	1518898000	473945856
1993	0.318211065	13180953598	1999349773	926303715
1994	0.356138727	16286433533	2428725001	1944515936
1995	0.388334428	20736164459	3108993366	1780400000
1996	0.455743274	24657470575	3742635769	2395000000
1997	0.584708333	26843700442	4425119856	2220000000
1998	0.608242811	27209602050	4665812621	1671000000
1999	0.603434888	28683659007	5075377296	1412000000
2000	0.671308552	31172518403	5750363582	1298000000
2001	0.757220692	32685198735	6425801618	1300000000
2002	0.868419917	35064105501	7173075035	1400000000
2003	0.957054376	39552513316	8040061270	1450000000
2004	1.090129494	45427854693	9238854312	1610000000
2005	1.170708252	57633255618	10848471132	1954000000
2006	1.214236115	66371664817	12863310252	2400000000
2007	1.227733963	77414425532	15003236547	6700000000
2008	1.368139953	99130304099	18418071125	9579000000
2009	1.476993533	1.06E+11	19401780477	7600000000
2010	1.622618922	1.16E+11	15008931942	8000000000
2011	1.712240646	1.36E+11	18100756957	7430000000
2012	1.583708122	1.56E+11	20700211254	8368000000
2013	1.622307629	1.71E+11	22832775311	8900000000
2014	1.819893977	1.86E+11	24539530925	9200000000
2015	N/A	1.93E+11	26463842087	11800000000
2016	N/A	2.05E+11	29283700778	12600000000
2017	N/A	2.24E+11	34308986384	14100000000
2018	N/A	2.45E+11	39225645461	15500000000

## C. Processes of predicting model

Since the values of CO<sub>2</sub>E from 2015 to 2018 are not available [14]. In order to predict the values, we proposed linear regression [17] based on observations from table I. Our algorithm examines different periods, K. Because our data is collected from World Bank [14] from 1986. Thus, the work examines three periods, corresponding K (5, 10, 15). As results, all CO<sub>2</sub>E values are completed replace with linear prediction, then we go to next step for multiple regression analysis. Intuitively, the proposed algorithm is not only

optimum replacement values of CO2E data from 2015 to 2018 but also accumulated growth rate of CO2E data.

The main idea behind the proposal help us to figure out how the different time periods will impact on dynamic change of CO2E in the future. Based on these results, advanced policies for each period will be recommended to attract either foreign investment in industrial section or not. Thus, our hope is to improve maximum chances to build a model economic to achieved rapid development and environment protection. Our main contribution is twofold 1) optimum replacement values of CO2E data, and 2) multiple regression analysis to show the relationships between CO2E and GDP, MAN, and FDI which is presented in Fig. 5.

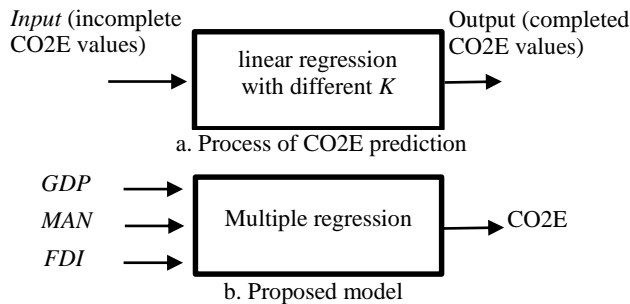


Fig. 5. Proposed process for multiple regression analysis

#### IV. RESULTS AND DISCUSSION

##### A. Results

In the first step, we perform linear regression [17] to predict the values for a time series data of CO2E from 2015 to 2018. The  $K$  values are selected for length of time series ( $K=5, 10, 15$ ). An iterative process for each  $K$  is presented until all CO2E values are replaced. As results, all in completed values of CO2E from 2015 to 2018 are completed and shown in table II. For all selected  $K$  sizes, though values of CO2E from 2015 to 2018 for all scenarios increase, each  $K$  present different fit line through the size of data. This refers that dynamic change of CO2E in the future depend on the policies on each period.

TABLE II. VALUES OF CO2E FROM 2015 TO 2018.

Year	K = 5	K = 10	K = 15
2015	1.763538987	1.935047562	1.830832243
2016	1.801972633	2.011980894	1.918060666
2017	1.835850887	2.087192423	2.003552097
2018	1.891462374	2.164784455	2.086096141

In the second step, we perform the multiple regression with dependent output CO2E for each  $K$ , independent inputs GDP, MAN, and FDI. The regression summary consists of three tables in which table III present the model summary, table IV shows coefficients, and table V is additional tests. To keep original results, we show all information of the regression results and focus on the most important sectors for identifying relationship among coefficients.

R-squared and Adj. R-squared evaluate the scatter of the data points around the fitted multiple regression line. Since R-squared values for each  $K$  (5, 15, 25) are 0.872, 0.918, and

0.905. The greater R-squared values, the smaller error between the observed values and expected values. In our model, F-statistic is 108.7 which follows an  $F$  distribution [18] and Prob (F-statistic) so that we can say the overall multiple regression is significant. All results for  $K = 15$  are presented in table III.

Table IV presents the coefficients of the intercept and the constant values for multiple regression. In addition, the other coefficients such as standard error (std err),  $t$  statistic model and its  $P$  value are presented. Standard errors of the coefficients we will calculate the covariance-variance factor,  $t$  statistic is given by the ratio of the coefficient (or factor) of the predictor variable of interest, and its corresponding standard error. The confidence interval is the range of values we would expect to find the parameter of interest and a smaller confidence interval suggests that we are confident about the value of the estimated coefficient. In the table IV, three of the coefficients are significant at the 5% level, and  $std\ err$  is the standard deviation of its sampling  $F$  distribution.

TABLE III. MODEL SUMMARY OF OLS REGRESSION

Dep. Variable:	CO2E	R-squared:	0.918
Model:	OLS	Adj. R-squared:	0.910
Method:	Least Squares	F-statistic:	108.7
Date:	Tue, 02 Jun 2020	Prob (F-statistic):	7.11e-16
Time:	15:07:23	Log-Likelihood:	10.412
No. Observations:	33	AIC:	-12.82
Df Residuals:	29	BIC:	-6.838
Df Model:	3		
Covariance Type:	nonrobust		

TABLE IV. THE COEFFICIENTS OF OLS REGRESSION

	coef	std err	t	P> t	[0.025	0.975]
const	1.0000	0.033	30.512	0.000	0.933	1.067
GDP	0.2349	0.200	1.175	0.250	-0.174	0.644
MAN	0.3265	0.191	1.713	0.097	-0.063	0.716
FDI	0.0337	0.130	0.260	0.797	-0.232	0.299

TABLE V. ADDITIONAL TESTS IN OLS REGRESSION

Omnibus:	0.544	Durbin-Watson:	0.294
Prob(Omnibus):	0.762	Jarque-Bera (JB):	0.666
Skew:	0.200	Prob(JB):	0.717
Kurtosis:	2.430	Cond. No.	13.6

Table V provides some additional information about the residuals of the model such as Omnibus, Skewness, Kurtosis, Durbin-Watson, Jarque-Bera (JB), and Cond. No in which Skew and kurtosis refer to the shape of a distribution, Omnibus test uses skewness and kurtosis to test the null hypothesis that a distribution is normal, The Durbin-Watson test is used to detect the presence of autocorrelation, and Jarque-Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

## B. Discussion

By using OLS [18], we find the parameter values for the predictive multiple regression equation in which constant value is 1, the other coefficient values for GDP, MAN, and FDI are corresponding 0.2349, 0.3265, and 0.0337. Thus, the relationship can be modelled as follows:

$$\ln(\text{CO2E}) = 1 + 0.2349 \ln(\text{GDP}) + 0.3265 \ln(\text{MAN}) + 0.0337 \ln(\text{FDI}) \quad (5)$$

From the equation (5), if the CO2E will increase one unit, then the GDP is expected to raise 0.2349 unit, while MAN and FDI expected to raise 0.3265 unit and 0.0337 unit. Moreover, we can find CO2E target at the year 2030, then we control the dependent variables so that they satisfy a target reduction of CO2E the equation (5). To make the explain is clear, we also show prediction of CO2E based on linear regression which is presented in Fig.6.

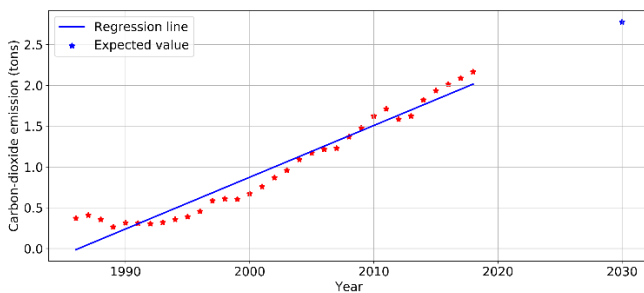


Fig. 6. The predicted CO2 emissions to 2030 in Vietnam

In table VI, we show the correlation coefficients of the variables. The correlation coefficients are used to measure the strength and direction of the linear relationship among CO2E, GDP, MAN, and FDI. Thus, Relationship of CO2E and MAN are stronger than relationship of CO2E and FDI, and relationship of CO2E and GDP. The correlation between two variables is helpful to determine how well a mutual fund performs relative to its benchmark index.

TABLE VI. THE CORRELATION COEFFICIENTS FOR CO2E, GDP, MAN, AND FDI

	CO2E	GDP	MAN	FDI
CO2E	1.000000	0.953208	0.955378	0.930073
GDP	0.953208	1.000000	0.984062	0.965260
MAN	0.955378	0.984062	1.000000	0.961701
FDI	0.930073	0.965260	0.961701	1.000000

## V. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we investigate and analyze relationship between carbon-dioxide emission and gross domestic

products, manufacturing sector and foreign direct investment. The empirical results reveal that there is positive relation among carbon dioxide emission and gross domestic products, manufacturing sector, and foreign direct investment. However, foreign direct investment and gross domestic products are weaker impact on carbon-dioxide emission.

The future work will investigate and analyze different manufacturing sections which are impact on carbon-dioxide emission. Thus, we can focus on how to pay attention on manufacturing sections which contribute smaller carbon-dioxide emission.

## REFERENCES

- [1] *The Fifth Ministerial Conference on Environment and Development*, 2005.
- [2] B. Trinh, K. Kobayashi, N. Q. Thai, N. V. Phong, and P. Le Hoa, "Analyzing some economic relations based on expansion input-output model," *International journal of business and management*, vol. 7, no. 19, p. 96, 2012.
- [3] G. M. Grossman and A. B. Krueger, "Environmental impacts of a North American free trade agreement," *National Bureau of Economic Research* 1991.
- [4] G. M. Grossman and A. B. J. T. q. j. o. e. Krueger, "Economic growth and the environment," vol. 110, no. 2, pp. 353-377, 1995.
- [5] C. Kiliç and F. J. P. Balan, "Is there an environmental Kuznets inverted-u shaped curve?," vol. 65, no. 1, pp. 79-94, 2018.
- [6] S. Dinda and D. J. E. E. Coondoo, "Income and emission: a panel data-based cointegration analysis," vol. 57, no. 2, pp. 167-181, 2006.
- [7] A. Jalil and M. J. E. E. Feridun, "The impact of growth, energy and financial development on the environment in China: a cointegration analysis," vol. 33, no. 2, pp. 284-291, 2011.
- [8] M. Nasir and F. U. J. E. P. Rehman, "Environmental Kuznets curve for carbon emissions in Pakistan: an empirical investigation," vol. 39, no. 3, pp. 1857-1864, 2011.
- [9] H.-T. Pao, H.-C. Yu, and Y.-H. J. E. Yang, "Modeling the CO2 emissions, energy use, and economic growth in Russia," vol. 36, no. 8, pp. 5094-5100, 2011.
- [10] A. G. Fisher, "Production, primary, secondary and tertiary," *Economic record*, vol. 15, no. 1, pp. 24-38, 1939.
- [11] T. Blasing and S. Jon, "Current greenhouse gas concentrations," *Updated February*, 2005.
- [12] E. Dlugokencky and P. Tans, "Trends in atmospheric carbon dioxide," vol. 25, 2013.
- [13] G. H. Brundtland, *Report of the World Commission on environment and development: "our common future"*. United Nations, 1987.
- [14] The World Bank data [Online]. Available: <http://www.worldbank.org/>
- [15] K. Saidi and S. J. E. R. Hammami, "The impact of CO2 emissions and economic growth on energy consumption in 58 countries," vol. 1, pp. 62-70, 2015.
- [16] P. S. Mann, *Introductory statistics*. John Wiley & Sons, 2007.
- [17] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019.
- [18] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *CVPR 2011*, 2011, pp. 785-792: IEEE.



# Proposed Research on Saline-Water Distillation for Living by Utilizing Waste Heat from Industrial Steam Boilers

Ngoc Han Pham

Faculty of Vehicle and Energy Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
hanngoc071193@gmail.com

Van Tuyen Nguyen

Faculty of Mechanical Engineering  
Ho Chi Minh City University of Technology  
Ho Chi Minh City, Vietnam  
tuyen55@gmail.com

**Abstract**— Demands on water as well as energy are raising rapidly with population increase; while fresh water in Vietnam is being a shortage, and industrial processes have exhausted a vast quantity of heat into environment. In this paper, a saline-water distillation system utilizing industrial boiler waste heat with low temperature water loop is proposed. This system assures the cleanliness of heat exchange surfaces. Furthermore, a condenser having two independent cooling water flows is used to guarantee the completely condensation of vapor in it to produce the fresh water and to heat the water for daily living. Calculation was relied on mass and energy balance equations, saline balance equations in the evaporator, and related equations on the physical properties of saline water, saturated vapor, and water. Case-study results for 10 ton/hour industrial steam boiler show that 208 kg/h of freshwater is collected and 2546 kg/h of running water can be heated respectively. With the chosen parameters, all of the heat exchangers equipped in the system are capable made in Vietnam.

**Keywords**— Fresh water; saline-water; industrial steam boiler; utilizing waste heat; water distillation.

## I. OVERVIEW ON WATER DISTILLATION UTILIZING WASTE HEAT

The fresh water scarcity is one of the most serious global challenges for our generation. Hence, searching for water production technologies is a priority mission. Over the years, the industry of fresh water production from seawater has developed powerfully; however energy consumption is on high-level. Relating to working principles, the processes of salt separation are classified in two main groups: thermal and membrane. The first one includes multi-stage flash distillation (MSF), multi-effect distillation (MED), vapor-compression distillation (VC), and freeze desalination (FD). The second group involves reverse osmosis (RO), membrane distillation (MD), and electrodialysis (ED).

Talking about thermal water distillation, some researches are carried out, using different sources of heat.

Roy et al. [1] have studied a Once-Through Multi-Stage-Flash (OT-MSF) desalination system, and learned about impact of top brine temperature (TBT) of up to 160°C on both the system design and performance characteristics. Mussati et al. [2] have considered the operation structures of MSF-mixer and OT-MSF systems; the results showed that the highest operating temperature of each stage is distinct, with value in the range of 90°C-120°C. Obviously, that is the factor effecting to the system thermal efficiency and leading to temperature difference between saltwater heater and condensing equipment. Ali et al. [3] have used Matlab

software to optimize operating parameters of the recirculating multi-stage flash desalting plant (MSF-BR).

Salimi et al. [4] have proposed the integration of multi-effect desalination system with a suitable power cycle for waste heat recovery. In this case, internal combustion engine (ICE) was chosen.

Hamed et al. [5] have conducted a dynamic analysis which is based on the first and second thermodynamic laws to evaluate the efficiency of the steam-heated desalination system. Chen et al. [6] have simulated the thermal vapor compression desalination process for high salinity water, using burning gas, to calculate the energy consumption and cost of product unit. The study showed that the cost of this simulated system is lowest compared to others.

In Vietnam, Liem [7] has applied the multi-effect water distillation method to produce freshwater using low-pressure steam extracted from thermal power plants which located in coastal areas. Phong [8] has assessed the ability to distill water by recovering latent heat contained in gas-turbine exhaust on Su Tu Trang oil rig.

Ana recently Bang [9] has studied the possibility of waste heat recovery from a rice husk fired boiler for distillation of fresh water from alum water. Results showed that all most 1 m<sup>3</sup>/h fresh water can be gained when recovering flue-gas heat from 18 t/h boiler. Nevertheless, there are some disadvantages to this work. Firstly, alum water will cause deposits inside tubes and then lower heat transfer. Secondly, the alum water flowing inside gas-water heat exchanger will cause unexpected damages. The last drawback, only small portion of heat removed from condenser is used for preheating the supplied alum water, while the huge left is exhausted out to environment.

The above-mentioned studies proclaimed that most of fresh water production systems are in large capacity, using high temperature sources for their good efficiency and desired economic effect. And key opportunities today are available in optimizing existing systems, developing technologies for chemically corrosive systems, and recovering low temperature waste heat. But some problems in this field still exist, as below:

1. Big systems for fresh-water production are often associated with other technologies for salvaging waste heat from these processes; so they become more complicated. A lot of equipment in the systems works under pressure, leading to high cost of making fresh water.

2. In small systems, the raw water is usually heated by waste heat directly. This design makes systems to be simple, but the water heater will work under hazardous condition because of deposit risk occurred on both tube sides. Moreover, a significant amount of heat contained in cooling water isn't utilized.

## II. PROPOSED WATER DISTILLATION SYSTEM

Based on results of the previous studies as well as by inheriting and developing from Bang' research [9], a novel water distillation system using exhaust-gas heat from industrial boiler is proposed, as showed in Figure 1.

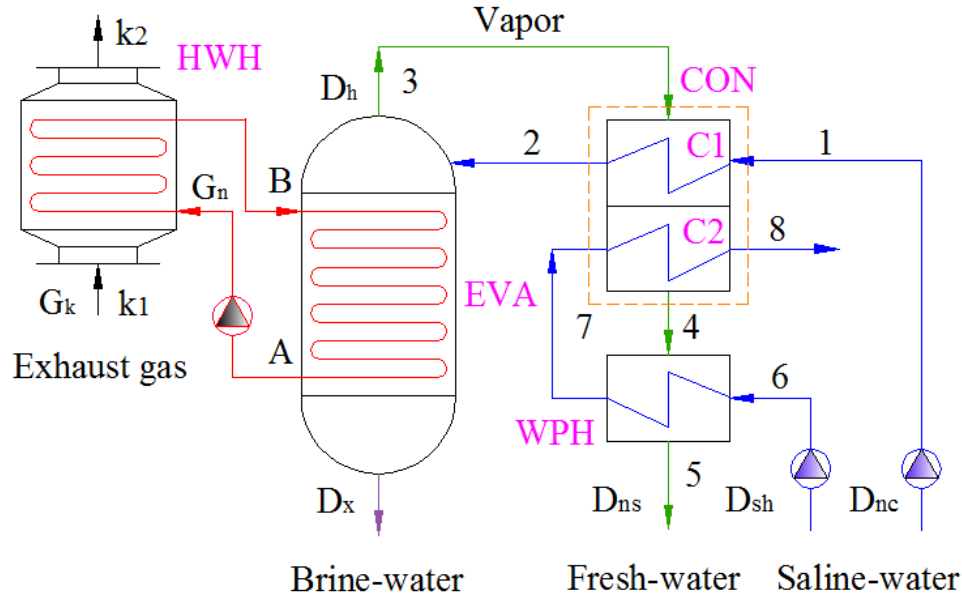


Fig. 1. Diagram of saline-water distillation for living utilizing waste heat.

The system has four (4) heat-exchangers: Gas-Water Heat-Exchanger (HWH), Evaporator (EVA), Condenser (CON), and Water Preheater (WPH). And three (3) pumps is used.

In heat exchanger HWH, boiler flue gas is used for heating clean water which runs around in a low temperature loop. This high-quality water ensures tube-inside cleanliness, leading to good heat transfer and to reduce O&M cost.

Owing to the heat received from the water loop, saline-water evaporation occurs in EVA, where vapor is produced. While a necessary amount of high-concentration brine water is likely to be blown down.

The condenser has two independent cooling water flows, which guarantee the entirely condensation of vapor in it. By this concept, the system can produce the freshwater and heat the daily living water as well.

Moreover, a heat exchanger is added right after the condenser to reduce temperature of produced fresh water and to preheat running water as the same time. So, a high waste-heat recovery efficiency of the proposed distillation system can be expected.

## III. WORKING PARAMETERS SELECTION

The main objective of the research is to bring fresh water out from saline-water as much as well to satisfy living demand in regions where domestic water is scarce. Consequently, available waste heat sources must be utilized maximum. On the other hand, system investment must be in an acceptable

limit. Therefore it is important to choose the suitable working parameters of the distillation system.

### A. Boiler exhaust-gas parameters

As a case study, exhaust gas from 10 ton/hour coal-fired industrial steam boiler is chosen as a potential waste heat resource. Assuming the boiler operates with 80 percent of designed steam load, its exhaust gas flow ( $G_k$ ) of 2 kg/s and temperature ( $t_{k1}$ ) of 240°C are typical. The waste gas will move in to the exchanger (HWH) for in-loop water heating.

Normally, final exhaust gas temperature ( $t_{k2}$ ) of 180°C right after a heat recovery device is expected. This exit

temperature is high enough against Acid Dew Point in order to avoid corrosion.

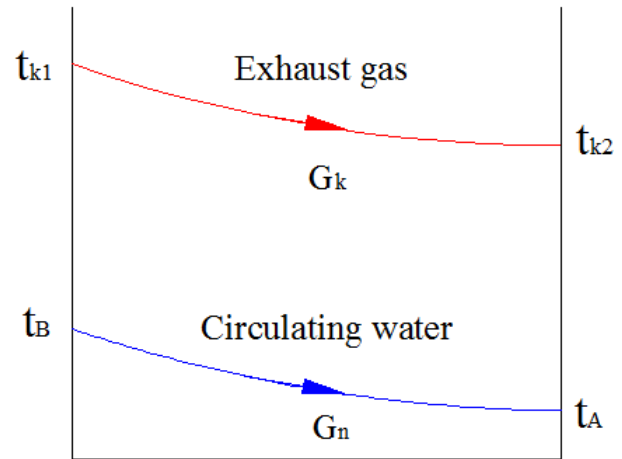


Fig. 2. Temperature profiles in Gas-Water Heat Exchange.

### B. Parameters of saline water

Vietnam is located in the region where sea-water has salinity concentration of less than 35000 ppm and average temperature of approximately 30°C. So in this study, the salt concentration ( $X_{nc}$ ) and temperature of the supply saline water ( $t_2$ ) might be selected according standards with 35 g/l and 30°C.

### C. Discharge rate of saline water

Acceptable discharge rate of saline water during distillation process has been studied previously. The total dissolved solids concentration in vapor as well as in the output produced water is assumed to be zero. The concentration of salinity in the discharge water ( $X_x$ ) is constrained at 70000 ppm as stated by industrial standards. The discharge rate of saline water ( $y$ ) is determined by the equilibrium salinity equation, and the calculated number is 0.5.

### D. Evaporator temperature

To get highest waste-heat recovery efficiency, the evaporate temperature must be as low as good. But in the other hand, the boiling pressure should be little higher than atmosphere for simply construction of the system. Therefore the boiling temperature ( $t_3$ ) of 105°C is selected.

### E. Temperature levels of circulating water

Suppose the lower temperature of in-loop water is about 3°C higher than boiling point in evaporator, and water temperature raising of 12°C takes place in the Gas-Water Heat Exchange. Thus the lower and higher temperatures of circulating water ( $t_A$  &  $t_B$ ) will get value 108°C & 120°C simultaneously.

### F. Parameters of condenser

Value of the condensing temperature is the same as boiling temperature ( $t_4 = t_3 = 105^\circ\text{C}$ ). There are two (2) independent cooling flows at the condenser.

At first, a certain portion of vapor condenses by supply saline-water in condenser part C1, which can be designed in order to raise water temperature ( $t_2$ ) up to 100°C. After that, the completely condensation takes place in part C2 by running water. The outlet temperature of domestic water is chosen about 60°C to meet common living demand.

Before entering in to condenser, the domestic water flow is used to cool the produced condensate down to chosen final temperature  $t_5 = 40^\circ\text{C}$  in Water Preheater (WPH).

Table 1 shows temperatures of fluids in specific points noted on above figure 1; except temperature at point 7 will be calculated later.

TABLE I. TEMPERATURES AT SELECTED POINTS

Parameters	Value, °C
Inlet exhaust gas temperature	240
Outlet exhaust gas temperature	180
Temperature at point A	108
Temperature at point B	120
Temperature at point 1	30
Temperature at point 2	100
Temperature at point 3	105
Temperature at point 4	105
Temperature at point 5	40
Temperature at point 6	30
Temperature at point 8	60

## IV. CALCULATING METHODS

The calculation is based on thermodynamic and physical properties of fluids, heat transfer equations, and mass equations.

### A. Physical properties of saline-water, saturated water and steam

According to literature [10], the following relations and equations were used. Here symbol “X” expresses salt concentration.

Density of saline-water is a function of salt concentration and temperature:

$$\rho_{nm} = 10^3(A_1F_1 + A_2F_2 + A_3F_3 + A_4F_4) \quad (1)$$

Where:

$$B = (2X - 150)/150, G_1 = 0.5, G_2 = B, G_3 = 2B_2 - 1$$

$$A_1 = 4.032G_1 + 0.115G_2 + 3.26 \times 10^{-4}G_3$$

$$A_2 = -0.108G_1 + 1.571 \times 10^{-3}G_2 - 4.23 \times 10^{-4}G_3$$

$$A_3 = -0.012G_1 + 1.74 \times 10^{-3}G_2 - 9 \times 10^{-6}G_3$$

$$A_4 = 6.92 \times 10^{-4}G_1 - 8.7 \times 10^{-5}G_2 - 5.3 \times 10^{-5}G_3$$

$$A = (2t - 200)/160, F_1 = 0.5, F_2 = A, F_3 = 2A^2 - 1, F_4 = 4A^3 - 3A$$

Thermal conductivity of saline-water,  $\lambda_{nm}$ :

$$\log_{10}(\lambda_{nm}) = \log_{10}(240 + 0.0002X) + 0.434 \left( 2.3 - \frac{343.5 + 0.037X}{t + 273.15} \right) \left( 1 - \frac{t + 273.15}{647 + 0.03X} \right)^{0.333} \quad (2)$$

Specific heat of saline-water at constant pressure:

$$c_{p,nm} = A + BT + CT^2 + DT^3 \quad (3)$$

Where:

$$A = 5.328 - 9.76 \times 10^{-2}X + 4.04 \times 10^{-4}X^2$$

$$B = -6.913 \times 10^{-3} + 7.351 \times 10^{-4}X - 3.15 \times 10^{-6}X^2$$

$$C = 9.6 \times 10^{-6} - 1.927 \times 10^{-6}X + 8.23 \times 10^{-9}X^2$$

$$D = 2.5 \times 10^{-9} + 1.666 \times 10^{-9}X - 7.125 \times 10^{-12}X^2$$

Dynamic viscosity of saline-water:

$$\mu_{nm} = \mu_w (1 + AX + BX^2) \quad (4)$$

Where:

$$A = 1.541 + 1.998 \times 10^{-2}t - 9.52 \times 10^{-5}t^2$$

$$B = 7.974 - 7.561 \times 10^{-2} t + 4.724 \times 10^{-4} t^2$$

$$\mu_w = 4.2844 \times 10^{-5} + \left( 0.157(t + 64.993)^2 - 91.296 \right)^{-1}$$

Surface tension of saline-water:

$$\sigma_{nm} = 0.2358 \left( 1 - \frac{t + 273.15}{647.096} \right)^{1.256} \left[ 1 - 0.625 \left( 1 - \frac{t + 273.15}{647.096} \right) \right] \quad (5)$$

Where:

Enthalpy of saturated vapor:

$$i'' = 2501.689845 + 1.806916015T + 5.087717 \times 10^{-4} T^2 - 1.1221 \times 10^{-5} T^3 \quad (6)$$

Enthalpy of saturated water:

$$i' = -0.033635409 + 4.207557011T - 6.200339 \times 10^{-4} T^2 + 4.459374 \times 10^{-6} T \quad (7)$$

#### B. Energy balance, mass balance, and heat transfer equations

Recovery heat in Gas-Water Heat-Exchanger (HWH), when heat loss into atmosphere is ignored:

$$Q_{HWH} = G_k c_{pk} (t_{k1} + t_{k2}) \quad (8)$$

Water flow in circulating loop:

$$D_n = \frac{Q_{HWH}}{c_{pn} (t_B - t_A)} \quad (9)$$

Heat balance for Evaporator:

$$Q_{EVA} = D_h i_3'' + D_x i_3' - D_{nc} i_2 \quad (10)$$

Mass balance for Evaporator:

$$D_{nc} = D_h + D_x \quad (11)$$

Salt concentration balance at Evaporator:

$$D_x = y D_{nc} \quad (12)$$

Heat balance for 1<sup>st</sup> part of Condenser, C1:

$$Q_{C1} = D_{C1} (i_3'' - i_4') = D_{nc} (i_2 - i_1) \quad (13)$$

Heat balance for 2<sup>nd</sup> part of Condenser, C2:

$$Q_{C2} = D_{C2} (i_3'' - i_4') = D_{sh} (i_8 - i_7) \quad (14)$$

Heat balance for Water Preheater (WPH):

$$Q_{WPH} = D_{sh} (i_7 - i_6) = D_{ns} (i_4' - i_5) \quad (15)$$

Heat exchanger areas are determined by heat transfer equation:

$$F_i = \frac{Q_i}{k_i \Delta T_i} \quad (16)$$

In which, i = [HWH, EVA, C1, C2, WPH] is respective at the heat exchanger, k is the heat exchange coefficient of the equipment and  $\Delta T$  is the logarithm average temperature difference.

Lastly, pump power is calculated:

$$W_j = \frac{D_j \Delta p_j}{\rho_j \eta_j} \quad (17)$$

Where, j = [n, sh, nc] corresponds the moving fluid, D is flow of the fluid,  $\Delta p$  is total pressure drop of particular flow. The local pressure drop of 0.5 bar by each tube bank and pump efficiency  $\eta = 0.8$  are predicted. Note that the water pressure in closed circulating loop must be high enough for boiling prevention.

#### C. Calculation flowchart

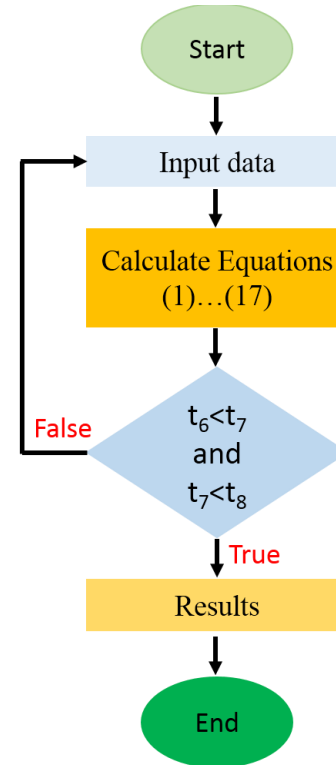


Fig. 3. Calculation flowchart.

A Matlab code program was written to solve the above set-up problem.

#### V. RESULTS AND DISCUSSION

Table II represents the main calculation results of the proposed distillation system using heat from 10 t/h steam boiler exhaust gas.

TABLE II. CALCULATED RESULTS OF THE PROPOSED SYSTEM

Parameters	Value	Units
Temperature at point 7	35.6	°C
Circulated water flow	9497	kg/h
Supplied water flow	417	kg/h
Domestic water flow	2546	kg/h
Produced clean water flow	208	kg/h
Brine-water flow	208	kg/h
Total of heat exchanger areas	20.8	m <sup>2</sup>
Total pumps power	448	W

Above received results show that the set-up distillation system using recovered heat from 10 t/h steam boiler can produce the amount of fresh water up to 208 kg/h (near 5 m<sup>3</sup> per day), and can raise temperature to 60°C for 2546 kg/h of domestic water (about 61 m<sup>3</sup> per day) concurrently. This amount of fresh water and hot domestic water can contribute to satisfactory in water-deficient areas. On the contrary, the electricity consumption is rather low. And benefits of waste heat recovery are clear in this case.

However, a significant saline-water blowdown was confirmed by calculation; and a remarkable heat obtained in this flow is wasted to environment. Obviously, additional energy saving should be obtained from the blowdown.

Based on the above case study, the calculation was expanded for larger distillation plants utilizing waste heat from industrial boilers with evaporation rate of 1 to 20 ton of steam per hour.

As seen on Figure 4, both feed saline water and fresh water flows increase with boiler output linearly. Using exhaust gas heat from boilers in the range from 1 to 20 ton of steam per hour, the amount of fresh water of 21 to 417 kg/h is achieved. And the supply water flow of 42 to 833 kg/h is required.

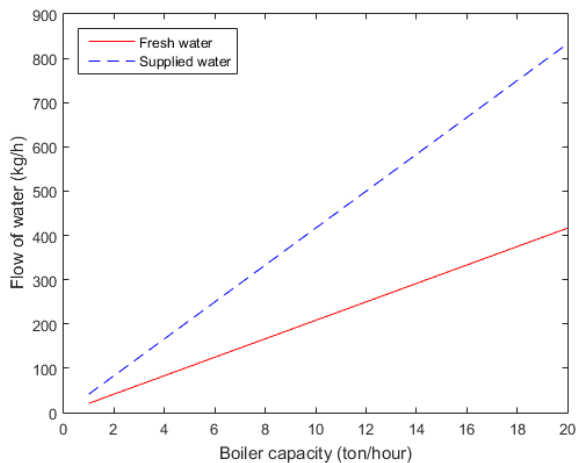


Fig. 4. Increase of supplied water and fresh water flows with boiler output.

There is a great number of domestic water heated in condenser, as shown in Figure 5, from 256 to 5091 kg/h according to boiler capacity. Hot domestic water flow is almost 6 times higher than supplied water flow. It proves that these distillation systems have ability to provide a huge amount of hot water to serve the daily life in the region by recovering boiler exhaust heat.

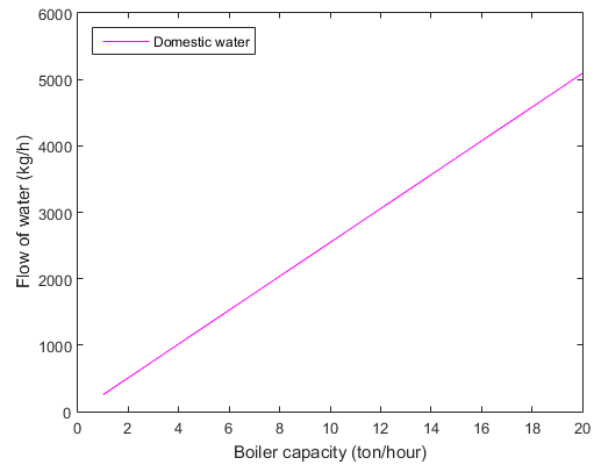


Fig. 5. Increase of domestic water flow with boiler output.

Obviously in order to produce fresh water, heat exchanges and water pumps must be installed as seen on the system diagram. Required total heat-exchanger areas and needed pump powers can be estimated as seen on Figure 6. Bigger distillation system asks higher investment as well as operating cost. Therefore, it is necessary to consider the size of designed distillation system to be appropriate to the particular waste heat source and living demands.

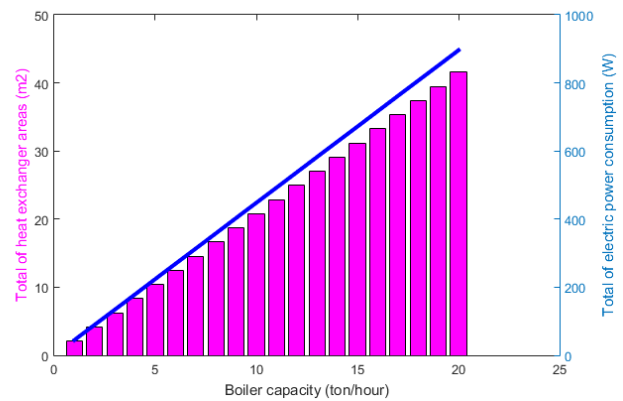


Fig. 6. Total heat-exchanger areas and total pump powers.

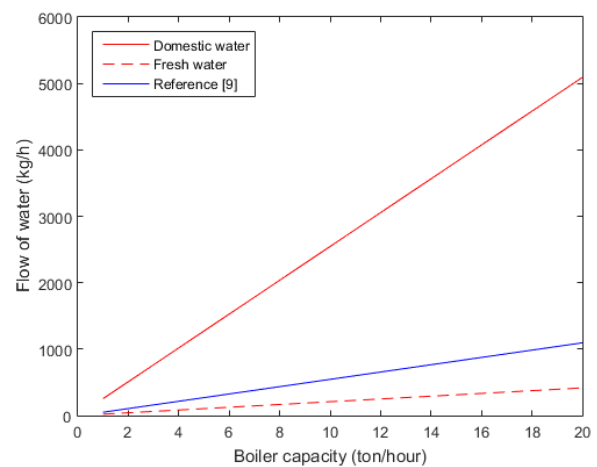


Fig. 7. Compare current proposed research with reference [9].



Figure 6 presents a comparison between the current proposed study and the reference [9]. This indicates that the amount of clean water produced in literature is higher than in this study. However, the proposed distillation system produces a sizable amount of potable water, about four times more than fresh water in literature. In parallel with the need of drinking water, all daily activities of people also depend on domestic water that this saline-water distillation system can adapt.

## VI. CONLUSSION

In this present work, the saline-water distillation system utilizing waste heat from boiler exhaust was proposed. The logical working parameters were chosen for satisfied solution. Calculated results for the case of 10 t/h steam boiler show that the amount of 208 kg/h drinking water is produced, and 2546 kg/h domestic water is heated up to 60°C at the same time. With the chosen parameters, all of the heat exchangers equipped in the system are capable made in Vietnam.

The results of this study might be considered as a basis for further designing, manufacturing, installing, testing, and then evaluating the proposed saline-water distillation system.

## REFERENCES

- [1] Yagnaseni Roy, Gregory P. Thiel, Mohamed A. Antar, John H. Lienhard V, "The effect of increased top brine temperature on the performance and design of OT-MSF using a case study," *Desalination*, vol. 412, 2017, pp. 32–38.
- [2] Sergio F. Mussati, Pio A. Aguirre, Nicolás J. Scenna, "Improving the efficiency of the MSF once through (MSF-OT) and MSF-mixer (MSF-M) evaporators," *Desalination*, vol. 166, 2004, pp. 141–151.
- [3] Mongi Ben Ali, Lakhdar Kairouani, "Multi-objective optimization of operating parameters of a MSF-BR desalination plant using solver optimization tool of Matlab software," *Desalination*, vol. 381, 2016, pp. 71–83.
- [4] Mohsen Salimi, Majid Amidpour, "Modeling, simulation, parametric study and economic assessment of reciprocating internal combustion engine integrated with multi-effect desalination unit," *Energy Conversion and Management*, vol. 138, 2017, pp. 299–311.
- [5] O. A. Hamed, A. M. Zamamiri, S. Aly and N. Lior, "Thermal performance and exergy analysis of a thermal vapor compression desalination system," *Energy Conversion Management*, vol. 37, 1996, pp. 379–387.
- [6] Liwen Chen, Qiang Xu, John L. Gossage, Helen H. Lou, "Simulation and economic evaluation of a coupled thermal vapor compression desalination process for produced water management," *Journal of Natural Gas Science and Engineering*, vol. 36, 2016, pp. 442–453.
- [7] Đinh Tiến Liêm, "Nghiên cứu khả năng sản xuất nước ngọt từ nước biển bằng phương pháp chưng cất đa hiệu ứng, sử dụng nguồn nhiệt trong nhà máy nhiệt điện Vĩnh Tân 2," *Đại học Bách Khoa Tp.Hồ Chí Minh*, 2013.
- [8] Thảm Trần Thanh Phong, "Nghiên cứu và thực nghiệm thiết bị chưng cất nước dạng thu hồi nhiệt ẩn ngưng tụ sử dụng nguồn nhiệt từ khói thải tuabin khí của các dàn khoan dầu khí ở Việt Nam," *Đại học Bách Khoa Tp.Hồ Chí Minh*, 2016.
- [9] Nguyễn Vũ Bằng, "Nghiên cứu khả năng sản xuất nước sạch từ nước phèn sử dụng nhiệt khói thải lò hơi," *Đại học Bách Khoa Tp.Hồ Chí Minh*, 2016.
- [10] Mostafa H. Sharqawy, John H. Lienhard V & Syed M. Zubair, "Thermophysical properties of seawater: a review of existing correlations and data," *Desalination and Water Treatment*, vol. 16, 2010, pp. 354–380.

# Surface Roughness Optimization for Grinding Parameters of SKS3 Steel on Cylindrical Grinding Machine

Thi-Minh Pham

Faculty of Mechanical Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
minhpt@bctech.edu.vn

Huy-Tuan Pham\*

Faculty of Mechanical Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
phtuan@hcmute.edu.vn

Van-Khien Nguyen

Faculty of Mechanical Engineering  
Nam Sai Gon Polytechnic College  
Ho Chi Minh City, Vietnam  
dongpho05@gmail.com

Quang-Khoa Dang

Faculty of Mechanical Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
khoa dq@hcmute.edu.vn

Duong Thi Van Anh

Faculty of Mechanical Engineering  
Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Vietnam  
anhdtv@hcmute.edu.vn

**Abstract**—Grinding is a machining method that could help to attain high precision grade and finest roughness grade. Due to such characteristics, grinding is often chosen as the final finishing method. The surface quality of a product depends on many factors including cutting depth, feed rate and rotation speed of the part. This paper will study the effect of these three machining parameters on the surface quality of SKS3 parts while machining on cylindrical grinding machines. Design of experiment using surface response model (RSM) will be applied to find the optimal machining parameters. Experimental design method Box-Behnken will be used to set up experimental runs. The input and output test data are processed by Minitab software to find the model of the regression equation before solving the optimization problem. The results of this study not only help to find the optimal grinding mode for SKS3 steel material, but it can also be extended to steels with similar properties and hardness.

**Keywords**—RSM (Response surface methodology), Box-Behnken, design of experiment.

## I. INTRODUCTION

The ever-developing science and technology has created a myriad of new materials for the manufacturing industry. This leads to the unceasing need to find the optimal machining parameters for these materials with specific machining methods, especially for hard and complex materials (i.e., ceramic, super alloys, titanium, etc.). Grinding is an indispensable task in the process of metal machining. In order to achieve high efficiency for grinding products, there are many factors that affect the manufacturing process by this method.

Nguyen Anh Tuan [1] studied the influences of the technical factors on the wear of grinding wheel and surface quality of part in profile grinding for the circular groove. These results were applied to grind the induction heating stainless steel materials of the internal rolling groove of the bearing 6208. Nguyen Tuan Linh [2] studied the multi-objective optimization of alloy steel grinding process on external circular grinding machine. Nguyen Tuan Nhan [3] studied the effect of some technological parameters on the surface roughness of carbon steel parts when machining on flat grinding machines.

In a nutshell, researches for grinding works are primarily focused on solving the best quality surface problems. In particular, surface quality depends on grinding materials, grinding machines and technological processing parameters. This paper studies the factors that have a great influence on surface quality and determine the optimal cutting mode when machining SKS3 steel on the JHU-2706H cylindrical grinding machine.

## II. EXPERIMENTAL PROCEDURES

There are many factors that may affect the surface quality of grinding process, including feed rate, work-piece speed, grinding wheel speed, depth of cut, grain size of grinding wheel, cooling lubrication condition [4] and dressing parameters of grinding wheel [5]. In this research, three machining parameters (depth of cut, feed rate and work-piece speed) are analyzed to investigate their effects on the surface roughness when grinding SKS3 steel on the cylindrical grinding machine. The chemical composition and physical properties of SKS3 steel are shown in Table I – II.

TABLE I. CHEMICAL COMPOSITION OF SKS3 STEEL

Component	C	Si	Mn	Cr	W	P	S
Content (weight %)	0.95	0.7	0.5	0.753	0.75	0.03	0.03
			–				
			1.0				

TABLE II. PHYSICAL PROPERTIES OF SKS3 STEEL.

Mechanical properties	Value
Pre-treated hardness (HB)	190 – 217
Heat treated hardness (HRC)	58 – 62
Yield Strength (hardened to 62HRC) (MPa)	2,200
Elongation (%)	14
Elastic Modulus (hardened to 62HRC) (GPa)	193
Density (kg/m <sup>3</sup> )	7,810

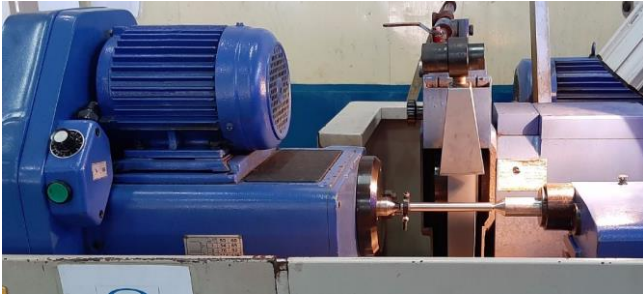


Fig. 1. Cylindrical grinding machine JHU-2706H

In this research, the cylindrical grinding machine (JHU-2706H, Jainnher Machine Co) and the grinding wheel WA60H (405 × 50 × 127) are used to grind SKS3 steel rods (Fig. 1). Dimensions of the work-pieces is (Ø15 × 151)mm. Fig. 2 is the instrument (Mitutoyo Surflest SJ-400, Japan) to measure the surface roughness.



Fig. 2. Surface roughness measuring instrument (Mitutoyo).

### III. DESIGN OF EXPERIMENTS

In this research, an experimental design technique based on a  $2^N$  factorial central composite-second-order rotatable design (CCD) [6] was proposed to use. This DoE technique has the advantages of reducing the size of experimentation but it still can improve the reliability of results without loss of accuracy. The purpose of the factorial experiments is to investigate the relationship between the response(s) and the levels of the design variables [7]. The number of experiments required by CCD method can be calculated by Eq. (1).

$$Q = 2^{N-f} + 2N + n_c \quad (1)$$

where  $Q, N, f, n_c$  are the total number of experiments, number of processing parameters, factorial number and number of replicates at the center point of the design space, respectively.

In this design,  $N = 3$  and  $f = 0$ . In order to do the analysis, only one experiment is necessary at the center of design space, therefore  $n_c = 1$ . From Eq. (1), there are 15 design points required for the DoE. The range of the three selected variables as input parameters are shown in Table III.

Three machining parameters ( $t, S_d$  and  $V_{ct}$ ) are coded into  $X_1, X_2$  and  $X_3$  using the formula in Eq. (2). Coded values of these independent variables in the experimental plan are shown in Table IV to conduct the experiments. Specimens after processing are measured in Fig. 3.

$$X_i = \frac{P_i - P_{i0}}{\Delta P_i} \quad (2)$$

where  $X_i$  is the coded value of the  $i^{th}$  independent variable,  $P_i$  the natural value of the  $i^{th}$  independent variable,  $P_{i0}$  the natural value of the  $i^{th}$  independent variable at the centre point, and  $\Delta P_i$  is the step change value.

TABLE III. FACTORS AND THEIR LEVELS

Input factors	Level			Unit
	Low	Centre	High	
Depth of cut ( $t$ )	0.002	0.01	0.02	mm
Feed rate ( $S_d$ )	3.2	4.0	4.8	mm/rpm
Work-piece speed ( $V_{ct}$ )	40	50	60	m/min

TABLE IV. BOX-BEHNKEN DESIGN OF EXPERIMENTS MATRIX

Exp. no.	$X_1$	$X_2$	$X_3$
1	-1	0	1
2	0	-1	1
3	1	0	1
4	0	1	1
5	-1	1	0
6	-1	-1	0
7	1	-1	0
8	1	1	0
9	-1	0	-1
10	0	-1	-1
11	1	0	-1
12	0	1	-1
13	-1	+1	+1
14	+1	-1	+1
15	0	0	0



Fig. 3. Fabricated specimens.

### IV. RESULTS AND DISCUSSION

#### A. Response surface modelling

To optimize the machining parameters of circular grinding for SKS3 steel in this research, RSM is used to design the layout of the experiments. RSM is the procedure for determining the relationship between various process parameters with the various machining criteria and exploring the effect of these process parameters on the coupled responses. The advantages of using the RSM method are that many factors can be optimized simultaneously and quantitative information can be obtained by only a few experimental trials [8].

In this research, a full second-order polynomial model is obtained by multiple regression technique for three factors by using analysis of variance (ANOVA). The regression model can be fitted into the following equation [6]:

$$Y = b_0 + \sum_{i=1}^k b_i X_i + \sum_{i=1}^k b_{ii} X_i^2 + \sum_{j>i}^k b_{ij} X_i X_j \quad (3)$$

Eq. (3) can be rewritten according to the three variables ( $X_1, X_2, X_3$ ) in the coded form:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{11}X_1^2 + b_{22}X_2^2 + b_{33}X_3^2 + b_{12}X_1X_2 + b_{13}X_1X_3 + b_{23}X_2X_3 \quad (4)$$

TABLE V. DESIGN OF EXPERIMENT AND THEIR RESULTS

Exp. no.	$t$ (mm)	$S_d$ (mm/rpm)	$V_{ct}$ (m/min)	$R_{atb}$
1	0.002	4.0	60	<b>0.152</b>
2	0.01	3.2	60	<b>0.155</b>
3	0.02	4.0	60	<b>0.142</b>
4	0.01	4.8	60	<b>0.155</b>
5	0.002	4.8	50	<b>0.157</b>
6	0.002	3.2	50	<b>1.114</b>
7	0.02	3.2	50	<b>1.111</b>
8	0.02	4.8	50	<b>1.052</b>
9	0.002	4.0	40	<b>0.975</b>
10	0.01	3.2	40	<b>0.923</b>
11	0.02	4.0	40	<b>1.035</b>
12	0.01	4.8	40	<b>0.979</b>
13	0.01	4.0	50	<b>0.975</b>
14	0.01	4.0	50	<b>0.930</b>
15	0.01	4.0	50	<b>0.900</b>

TABLE VI. ANOVA RESULTS

Term	Coef	SE Coef	T	P
Constant	-10,09	1,638	-6,159	0,002
t	179,21	29,850	6,004	0,002
Sd	4,22	0,517	8,161	0,000
Vct	0,10	0,041	2,358	0,065
t*t	-5447,22	463,690	-11,748	0,000
Sd*Sd	-0,49	0,058	-8,458	0,000
Vct*Vct	-9,2E-4	0,000	-2,491	0,055
t*Sd	1,64	4,921	0,332	0,753
t*Vct	-1,17	0,394	-2,973	0,031
Sd*Vct	-0,01	0,004	-1,449	0,207

S = 0,0710776 PRESS = 0,287247  
R-Sq = 98,70% R-Sq(pred) = 85,25%  
R-Sq(adj) = 96,37%

Using the results presented in Tables V, RSM is used to determine the regression mathematical model for  $R_a$ . The coefficients ( $b_i, b_{ii}$  and  $b_{ij}$ ) for the derived model are shown in Table VI.

All coefficients have sufficiently low p-value with  $R^2$  larger than 98%. Therefore they are significant in statistics. Also, the final models tested by variance analysis (F-test) indicated that the adequacy of models was established [9]. The form of the derived model is shown in Eq. (5).

$$R_a = -10.09 + 179.21t + 4.22S_d + 0.1V_{ct} - 5447.22t^2 - 0.49S_d^2 - 9.2 \times 10^{-4}V_{ct}^2 + 1.64tS_d - 1.17tV_{ct} - 0.01S_dV_{ct} \quad (5)$$

where  $X_1 = t, X_2 = S_d, X_3 = V_{ct}$

ANOVA analysis of the mathematical model shows the contribution level of each source to  $R_a$  is shown in Table VII. It is indicated that  $V_{ct}$  makes largest contribution with 43.99%. The contributions of  $t^2, S_d^2, t*V_{ct}$  and  $V_{ct}^2$  are 31.55%, 17.82%, 2.29% and 1.61%, respectively. In order to

optimize the surface roughness, it is suggested that  $V_{ct}$  could be targeted first. It is then followed by  $t^2$  and  $S_d^2$ .

TABLE VII. ANOVA ANALYSIS FOR  $R_a$ 

Source	dof	Contribution (%)	F-value	p-value
t	1	0.86	36.05	0.002
Sd	1	0	66.61	0.000
Vct	1	43.99	5.56	0.065
t*t	1	31.55	138.00	0.000
Sd*Sd	1	17.82	71.54	0.000
Vct*Vct	1	1.61	6.20	0.055
t*Sd	1	0.03	0.11	0.753
t*Vct	1	2.29	8.84	0.031
Sd*Vct	1	0.54	2.10	0.207
Error	5	1.29		
Total	14	100		

### B. Effect of working parameters on the surface roughness

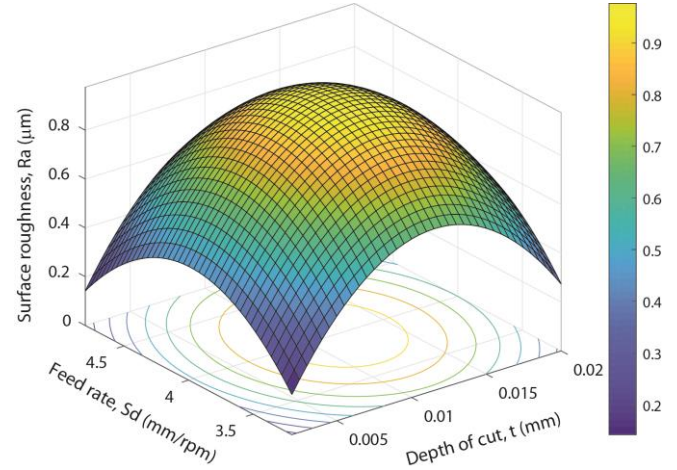


Fig. 4. Effect of depth of cut and feed rate on the surface roughness

Fig. 4 – 6 show the effect of different working parameters on the surface roughness of the SKS3 grinding specimens. It can be seen in Fig. 4 that when depth of cut and feed rate are increased to the middle range of the processed values, the surface roughness is largest or worst in term of surface quality. However, keep increasing these two parameters helps to improve the surface roughness of the machined products.

In Fig. 5, the positive proportion of both depth of cut and work-piece speed leads to the initial increment of the surface roughness when they are surging. However, the effect of depth of cut is the same as Fig. 4 when  $t$  is in the middle of the processing range. Surface roughness is largest at this point and is descending when  $t$  is increasing. On the contrary,  $R_a$  is monotonically ascending with  $V_{ct}$ .

The effect of feed rate and work-piece speed on the surface roughness in Fig. 6 is also the same as the effect of depth of cut and work-piece speed in Fig. 5. The surface roughness is largest in the middle range of feed rate and it is monotonically ascending with  $V_{ct}$ .

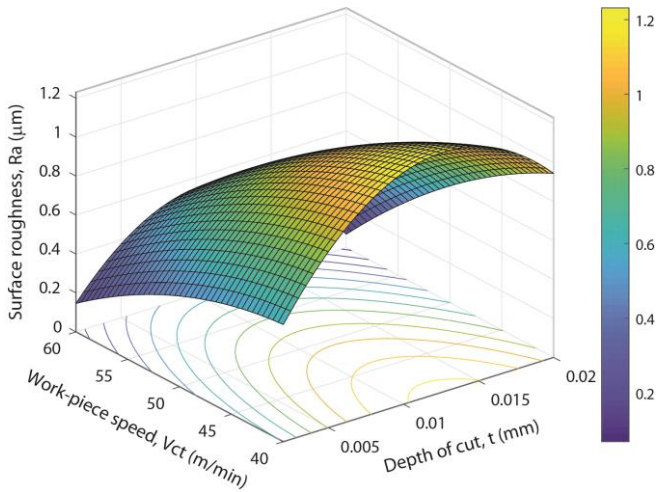


Fig. 5. Effect of depth of cut and work-piece speed on the surface roughness

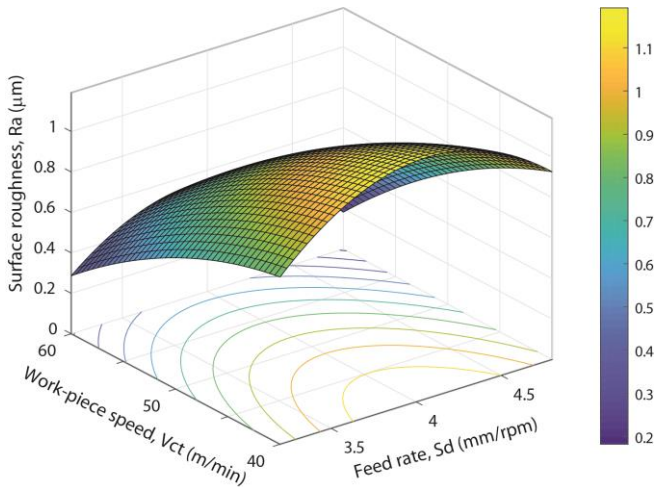


Fig. 6. Effect of feed rate and work-piece speed on the surface roughness

### C. Optimization

In order to find the optimum manufacturing condition for grinding SKS3 steel, this section will optimize Eq. (5) following the formulation in Table VIII.

TABLE VIII. OPTIMIZATION FORMULATION FOR SURFACE ROUGHNESS

• Objective function: Minimize Ra in Eq. (5)	
• Design variables: $t$ , $S_d$ and $V_{ct}$	
• Constraints:	
$0.002 \leq t \leq 0.02$	(6)
$3.2 \leq S_d \leq 4.8$	(7)
$40 \leq V_{ct} \leq 60$	(8)

### Response Optimization

#### Parameters

Goal	Lower	Target	Upper	Weight	Import
Ra	Target	0,07	0,1	0,5	1

#### Global Solution

$t$	=	0,002
$S_d$	=	3,2
$V_{ct}$	=	53,1313

#### Predicted Responses

Ra	=	0,100139,	desirability =	0,999653
----	---	-----------	----------------	----------

Composite Desirability = 0,999653

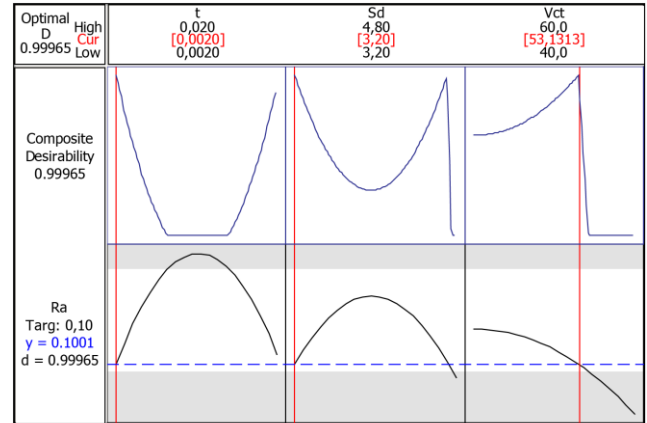


Fig. 7. Optimization results.

Using RSM method and minitab to optimize the surface roughness regression model, the optimization results are shown in Fig. 7. The independent effect of three processing parameters are shown separately in three columns from low to high level. The optimum values following a specific objective function is shown in the middle of low and high (Cur). For instant, if the targeted surface roughness is set at  $0.1\mu\text{m}$  with the lowest value  $0.05\mu\text{m}$  and the highest value  $0.5\mu\text{m}$ , the optimum results with the composite desirability  $d=0.999653$  for ( $t$ ,  $S_d$  and  $V_{ct}$ ) are found at  $0.002\text{mm}$ ,  $3.2\text{mm/rpm}$ , and  $53.1313\text{m/min}$ , respectively.

These optimum results found by RSM method can be used as referenced processing parameters when machining SKS3 steel or any other metal alloys with similar physical properties to increase the productivity of the manufacturing processes.

### V. CONCLUSIONS

The goal of this paper is to find the optimal parameters of the cutting mode that affects surface quality when grinding SKS3 steel on cylindrical grinding machines. This paper used RSM method and minitab software to find the regression coefficients of the mathematical model. Optimization is also implemented to the regression model to find the appropriate processing parameters with a targeted surface roughness. Effects of each parameters to the surface response are also analyzed. These results not only allow determining the optimal cutting mode when processing SKS3 steel but they can be also applied to other metal alloys with similar hardness such as NAK 40 steel.



# REFERENCES

- [1] Nguyen Anh Tuan, "Study on the influences of the technical factors on the wear of grinding wheel and the surface quality of part in profile grinding for the circular groove," PhD. dissertation, Hanoi University of Science and Technology, 2018.
- [2] Nguyen Tuan Linh, "Multi-objective optimization of grinding alloy steel on external cylindrical grinding machine," PhD. dissertation, Hanoi University of Science and Technology, 2015.
- [3] Nguyen Tuan Nhan, "The effect of some technological parameters on the surface roughness of carbon steel parts when machining on flat grinding machines," Master thesis, The University of Da Nang, 2014.
- [4] Khan, A. M. and et al., "Multi-objective optimization for grinding of AISI D2 steel with Al<sub>2</sub>O<sub>3</sub> wheel under MQL," *Materials*, vol. 11 (2018), pp. 2269.
- [5] Palmer, J., et al., "An experimental study of the effects of dressing parameters on the topography of grinding wheels during roller dressing," *Journal of Manufacturing Processes*, vol. 31 (2018), pp. 348–355.
- [6] Box G.E. and Hunter J.S., "Multifactor experimental designs," *Ann. Math. Stat.* (1957), pp. 28–195.
- [7] Hewidy, M. S., et al., "Modelling the machining parameters of wire electrical discharge machining of Inconel 601 using RSM," *Journal of Materials Processing Technology*, vol. 169 (2005), pp. 328–336.
- [8] Myers R.H., Montgomery D.C., and Anderson-Cook C.M., *Response Surface Methodology – Process and Product Optimization using Designed Experiments*, Wiley, 2016.
- [9] Montgomery D. C., Runger G.C., *Applied Statistics and Probability for Engineers*, 7<sup>th</sup> Ed., Wiley, 2018.

# Study on INTOC Waterproofing Technology for Basement of High-Rise Buildings

Sy-Hung Nguyen  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
sihung.nguyen@hcmute.edu.vn

Thanh-Tich Do  
INTOC Company  
Ho Chi Minh City, Vietnam  
thanhtich3699@gmail.com

Julien Ambre  
Université Côte d'Azur, CNRS,  
Observatoire de la Côte d'Azur, IRD,  
Géoazur, 06560 Valbonne, France  
ambre@geoazur.unice.fr

**Abstract**—The seepage or water leaks in general and the seepage of the basement of high-rise buildings are common problems in the world. In Vietnam, for more than 20 years, INTOC products have shown high efficiency in basement waterproofing through surveys of several constructions. However, research on these products is hardly available. In this study, the author conducted a number of experiments on an INTOC product, positive and negative sides' waterproofing effects as well as the influence of the product on the adhesion between reinforcement and concrete. The results showed that when applying the product on the concrete samples of grade M300 and M400 at water pressures of 10 and 12 daN/cm<sup>2</sup>, there wasn't any seepage flow through the waterproofing layer. The waterproofing slurry on the reinforced surface didn't affect its adhesion with concrete. The waterproofing layer has a very small surface tension, so it is hydrophobic and resists capillary permeability, and hence has small wetting properties.

**Keywords**—Waterproof, basement, seepage, permeability

## I. INTRODUCTION

### A. Overview of waterproofing

In the world as well as Vietnam, seepage in the buildings is very common. According to Nguyen D.H [1], there was 84.35% of construction under 10 years old in Ho Chi Minh city which seeped, of which the percentage of the basement was 78.3%, of which 15.7% seeped severely. According to Wader, J. [2], the rates of seepage constructions infiltrated in the US and Singapore were 60% and 53%, respectively.

Thus, seepage in building constructions is a difficult problem to resolve which causes many consequences for quality, performance and the long-term durability of buildings due to deterioration of the concrete. Moreover, the main type of seepage in basements is the negative-side. This leads to expensive structural repairing work, operational downtime, damage or loss of interior finishes and goods, and many other consequences.

In Vietnam, particularly in Ho Chi Minh City, the causes of seepage can include:

- A weak clay surface layer with a high water table which leads to high water pressure on the basement wall.
- Poor construction design;
- Water prevention and waterproofing measures during the construction phase which were not inappropriate;
- Concrete and waterproofing materials of poor quality.

- Cracks that could be generated in concrete walls or basement floors due to shrinkage, or excessive settlement of the building.

### B. Negative-side waterproofing methods

Positive-side waterproofing is applied to the wet or exterior side of foundations or slabs on grade and below grade, as well as suspended slabs. It is the predominant type of waterproofing used in new constructions. However, positive-side waterproofing for basements is difficult to apply for basement walls because of the large depth of the excavated pit.

Negative-side waterproofing is applied to the dry or inside side of the subsurface. It is used primarily for water-holding purposes. Negative-side waterproofing prevents water from entering occupied space. For example, waterproofing inside basement walls, outside water lakes (without draining water), or footwall waterproofing are considered negative-side ones. In particular, negative-side waterproofing for basement walls with high water pressure is considered the biggest challenge. Meanwhile, it is difficult to implement thoroughly, even considered impossible for developed countries with advanced construction levels such as the US [3], [4].

A common solution in dealing with negative-side seepage is to establish an interior perimeter drainage system around and inside of the basement walls. The system includes ditches, water pumps, dehumidifiers, and shielding auxiliary walls (Figure 1). The ditches collect water seeped from the basement walls, then pump it out. To ensure aesthetics and safety for the users, an additional exterior wall is also needed for covering. The dehumidifiers must be in non-stop operation.



Fig. 1. (a) A schematic of basement-drainage system. (b) a drainage system on construction drain. 1: channel drain (ditch), 2: water pump, 3: Barrette wall

The interior drainage system with ditches is commonly used in the world. In many constructions, this system is proactively built from the beginning. Research by the

University of Minnesota - USA confirms that the drainage system is the most effective method of waterproofing for basements, also considered mandatory to keep moisture away [5].

However, the drainage-system solution has limitations such as the requirement to maintain the system for at least 50 years of the project's life, causing moldy conditions, narrowing basement area, and high cost.

Another solution for basement waterproofing is using membranes or coating inside. This method is less expensive than a drainage system and seems to work for a limited time in some cases. However, the water is still there and eventually these systems deteriorate or simply move the water to another pathway into the basement. In addition, this solution may temporarily cover the leak, but when the water pressure from the outside ground is large enough, the water will enter the basement [5].

### C. Water-permeability of concrete

The permeability of concrete depends on both concrete and the viscosity of liquids. The Hagen-Poiseuille equation applied to non-compressible liquids and permanent flow based on Darcy's basic equation is:

$$k = \frac{\mu L Q}{A \Delta P} \quad (1)$$

Where k: coefficient of permeability in cm/sec,  $\mu$ : liquid viscosity (Pa.s),  $\Delta P$ : water pressure in kg/cm<sup>2</sup>, L: length of test specimen in cm, A: cross-sectional area of test specimen in cm<sup>2</sup>, Q: net rate of inflow in cm<sup>3</sup>/sec

As for concrete, the intrinsic factors like porosity, the sizes of the pores, the roughness of the cavity walls, the tortuosity and the connectivity of the voids [6]. The permeability coefficient will exponentially increase when the porosity is greater than 25% [7]. The porosity coefficient of concrete depends on the concrete mix, the ratio of water/cement and the aggregate size. As the ratio of water/cement increases (fluctuating in the range of 0.27 to 0.4), the mortar becomes looser and tends to penetrate aggregate crevices more easily, thereby reduces interconnection between pores and the permeability of concrete. However, when the ratio of water/cement is higher than 0.4, the permeability coefficient will increase significantly with the increase in the ratio of water/cement due to excess water evaporation creating many pores in concrete [8]. The bigger the aggregate size of concrete, the higher the permeability, due to the more microcracks that develop around the larger aggregate [9]. In addition, the permeability of concrete depends on additives and fine fillers such as fly-ash [10,11,12]. The permeability of reinforced concrete also depends on the diameter and the density of reinforcement, and the thickness of the protective concrete layer [13,14].

Many studies show the dependence between the water permeability coefficient of concrete and concrete grade. The higher the compressive grade concrete, the smaller the permeability coefficient. For example, in an EDF-France study, it was shown that with C25/C30 or higher concrete grade, the permeability coefficient became smaller and stable in the range of 10-11m/s. [15]. In addition, the permeability of concrete depends on the nature and the magnitude of cracks and stress states [16].

### D. INTOC waterproofing products

INTOC waterproofing products have been put on the market for more than 20 years and successfully processed many items with high water pressure such as lakes, overhead swimming pools, lift pits ... [17, 18, 19], especially basements with negative seepage problems without drainage systems.

INTOC waterproofing products studied here carry the code of INTOC-04 in liquid form with components including extracted cement, quartz powder and improved silicate combined with water-resistant compounds. INTOC-04 is mixed with cement and water in a certain proportion to create a waterproof slurry. When the slurry is poured on the concrete surface, it will penetrate the concrete thanks to the capillary mechanism and seal the pores. On the other hand, the slurry hardens and forms a film covering on the concrete surface with a certain thickness and very small permeability coefficient. In addition, the waterproof membrane has a very small attraction with water molecules [20].

INTOC waterproofing layers do not deteriorate over time like organic-based ones. Moreover, due to its cement base, the adhesion between the waterproofing layer and the concrete is very strong. INTOC products can be used for both positive- and negative-side waterproofing.

## II. INTOC WATERPROOFING TECHNOLOGY

INTOC-04 has a density of 1.05kg/litre and a pH degree in the range of 8 to 10. When mixing INTOC-04 with a proportion of 1kg INTOC-04 + 3kg water + 8kg of cement, a waterproofing slurry is formed. After 24 hours, the slurry will be solidified and reach the level of waterproofing required [20].

### A. Negative-side waterproofing process for basement walls

INTOC's waterproofing process consists of the following steps:

#### 1) Preparation

The surface of the walls must be rough enough to ensure good adhesion between the concrete and the waterproofing layer. If the concrete surfaces are smooth, the roughness is caused by slanted cuts to them. The cuts into the concrete surfaces have an angle of 45 degrees with 1.5cm of depth and 10cm of length. The distance between the two cutting lines is about 15cm. The cutting lines are arranged alternatively, vertically and horizontally (Figure 2). The back rows are inclined in the opposite direction to the front ones.

#### 2) Plastic pipe attachment

For the high efficiency of negative-side waterproofing, it is necessary to attach the plastic-pipe method (Figure 2). This method will reduce the water pressure, ensuring that the waterproofing layers do not drift by the seepage flow. Drilling a small hole with 5cm of depth into the wall, attaching and fixing the plastic pipes about 2-3cm deep down the hole. Thus, the water seeped concentrates and flows out along the pipes.

#### 3) Coating of waterproofing slurry

After attaching the pipes, the wall is made damped to the water saturation. Then the waterproof slurry is coated on the surface of the wall. After at least 24 hours, when the waterproofing slurry hardens, a protective mortar layer with 10mm of thickness covers above the waterproofing layer.

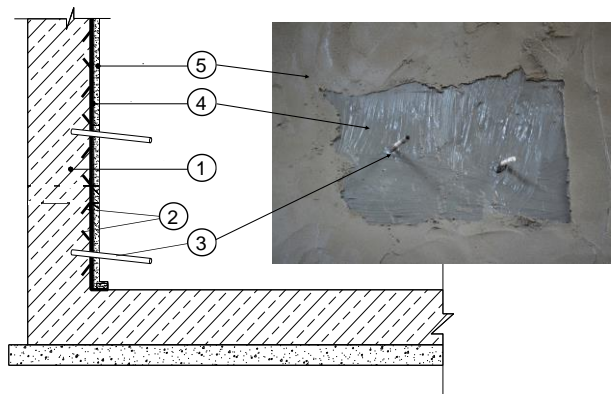


Fig. 2. A schematic of waterproofing method of INTOC. 1: concrete wall , 2: cuts of 450, 3: plastic pipes, 4: waterproofing layer, 5: protective mortar

### B. Positive-side waterproofing process for basement floors

After finishing the lean concrete layer and installing the reinforcement for basement floors, the waterproofing slurry is poured on the lean concrete surface. Finally, the concrete floor is poured on the lean concrete with a waterproofing layer. Thus, the waterproofing layer will adhere to the bottom of the basement floors, and not be separated from the floor concrete.

### C. Surveying on some constructions treated by INTOC-04 waterproofing product

The authors surveyed a number of constructions in Ho Chi Minh city – Viet Nam and a construction in Myanmar which were treated by INTOC-04 waterproofing product. The information is gathered in Table I.

The majority of the constructions with barrette basement walls listed in the table 1 are in Ho Chi Minh city, which has a high groundwater level. The walls often showed seepage problems right during the construction process. For all constructions, the seepage took place on a large scale, even at some locations with strong water flows. The wall concrete was often of poor quality, with many holes, cavities or/and cracks, many of whose reinforcement bars were exposed (Figure 3). In addition, the water could flow through torn panel joints. The causes of poor concrete quality in barrette walls were as follows:

- Complicated geological conditions with a high level of groundwater and poor supporting fluid (drilling fluids such as bentonite or polymers) quality. As a result, the soil could be expected to collapse during the concreting of the walls and the concrete could be contaminated.
- Lack of stability and workability of tremie concrete, which could lead to typical damages on walls like bleeding channels and honeycombs (lack of stability) as well as an insufficient concrete cover, bond with reinforcement and leaks (lack of workability);
- Limited availability and capacity of equipment;
- Lack of skills in equipment operation.

The constructions mentioned above, after being repaired by INTOC-04 waterproofing product with INTOC technology, haven't shown any seepage problem until now. All constructions in the survey are still of good quality. The construction with the longest time under waterproofing has been under waterproofing for 13 years.

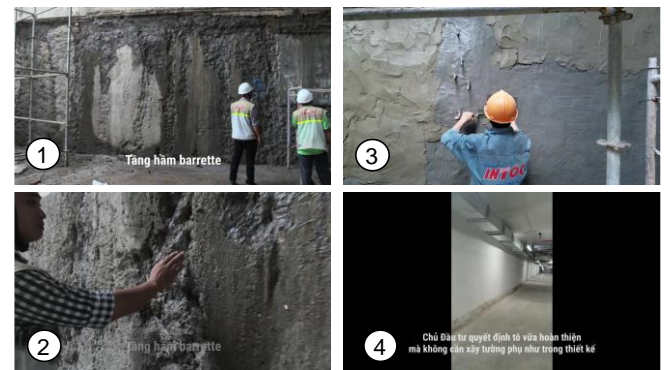


Fig. 3. Waterproofing process at Gigamall Centre – HCM city. 1,2: Barrette wall with seepage before waterproofing, 3: waterproofing execution, 4: Wall after waterproofing

TABLE I. CONSTRUCTIONS TREATED BY INTOC-04 WATERPROOFING PRODUCT

N <sup>o</sup>	Construction type, address	Seeped items	Condition, seepage level	Remedial waterproofing solution	Amount of time under treatment until 03/2020 (year)
1	Giant Hypermarket - Gigamall Centre – HCM city	Barrette walls of basement with a thickness of 800 mm	The basement has two floors. Barrette walls seeped right after construction. The seepage level is severe on a large scale.	Negative-side waterproofing for barrette walls	07/2018
2	Hospital 115 – HCM city	Barrette walls of basement with a thickness of 600 mm	Barrette walls seeped right after construction. The seepage level is severe on a large scale.	Negative-side waterproofing for barrette walls	10/2018
3	Hoang Anh Gia Lai Complex (phase 1) - Yangon, Myanmar	Concrete walls of basement with a total area of 50.000 m <sup>2</sup>	There is a big lake close to Complex (about 50m) with a high level of water. The seepage level is severe on a large scale. There is strong water flow leaking through holes and cracks on the walls.	Negative-side waterproofing for barrette walls	2014
4	Moonlight Park View Apartment – HCM city	Barrette walls of basement with a thickness of 600 mm	Barrette walls seeped right after construction. The seepage level is severe on a large scale. There are several defects of wall concrete such as holes and cracks.	Negative-side waterproofing for barrette walls	2018

N <sup>o</sup>	Construction type, address	Seeped items	Condition, seepage level	Remedial waterproofing solution	Amount of time under treatment until 03/2020 (year)
5	Tan Phu Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm		Negative-side waterproofing for barrette walls	2007
6	Hoc Mon Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm		Negative-side waterproofing for barrette walls	2007
7	Go Vap Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm		Negative-side waterproofing for barrette walls	2009
8	HCM city Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm	The basement has two floors Barrette walls seeped right after construction. The seepage level is severe on a large scale.	Negative-side waterproofing for barrette walls	2010
9	Gia Dinh Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm		Negative-side waterproofing for barrette walls	2011
10	Duyen Hai Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm		Negative-side waterproofing for barrette walls	2011
11	Tan Binh Power company headquarters – HCM city	Barrette walls of basement with a thickness of 600 mm		Negative-side waterproofing for barrette walls	2011

### III. TESTING PROGRAM

#### A. Experiment on surface tension between water and waterproofing layer

Surface tension is the tendency of liquid surfaces to shrink into the minimum surface area possible. Surface tension is an important factor in capillarity. Beading of water on a solid surface, if there is an attraction between the liquid and the solid, the drop placed on the solid will tend to spread; If there is repulsion between the liquid and the solid, the drop placed on the solid will tend to “regroup”, to take a spherical shape; Water adheres weakly to solid and strongly to itself, so water clusters into drops. Surface tension gives them their near-spherical shape because a sphere has the smallest possible surface area to volume ratio.

In this study, the experiments were conducted as follows:

- Manufacturing a waterproofing sample with a thickness of approximately 4 mm, a diameter of 20 cm.
- After 24 hours, dropping the water droplets on the surface of the sample and observing the shape of the droplets.
- Breaking the sample into two halves, then combining them so that the gap between them is approximately 0.3mm wide. Placing the two halves on a piece of absorbent paper. Putting water droplets into the gap; after 1 minute, observing the absorbent paper to see if water penetrates the gap;
- Doing the same test for a sample of cement-sand M75 at 28 days for comparison.

Experimental results show that (Figure 4):

When dripping water on the surface of the waterproofing sample, the shape of the droplets is a hemisphere which does not sticky spread to the surface of the sample. Thus, the waterproofing sample is a hydrophobic material. When dripping water into the gap between the two halves of the sample, after 1 minute, no water was detected on the surface of the absorbent paper and water droplets on the surface are not widespread. Thus, due to hydrophobicity, water was unable to penetrate the gap between the two half samples.

Doing similar experiments on cement-sand samples, the water droplets dripped on the surface were widespread and not have a hemispherical shape. The water droplets dripped on the gap of the halves penetrated easily through the interstitial.

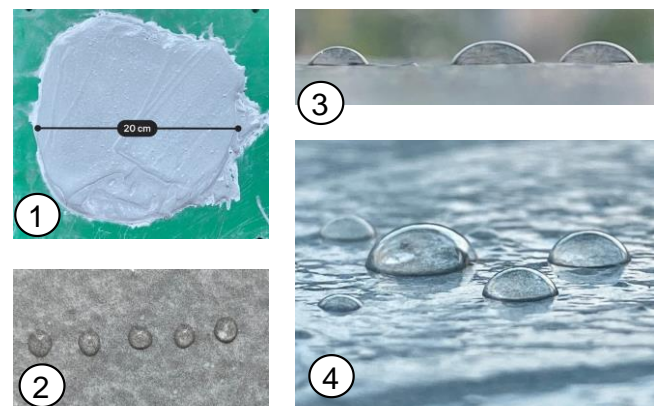


Fig. 4. Tension test of water on the waterproofing surface layer. 1: waterproofing sample, 2: view of droplets from above, 3: Frontal view, 4: 3D view



### B. Experiments of positive-side and negative-side waterproofing effect

Tests of waterproofing including both positive- and negative-side seepage were conducted on cylindrical concrete samples with a diameter of 150 mm and a height of 150 mm, with two concrete grades, M300 and M400. The concrete mixtures are shown in Table II.

TABLE II. MIXTURES OF CONCRETE (FOR 1 M3 FRESH CONCRETE)

Material	M300	M400
Cement (kg)	365	500
Sand (kg)	760	630
Crushed stone 1x2 (kg)	1060	1065
Water (lit)	205	205

The experimental process is prepared with the following steps;

- Casting cylindrical specimens with 150mm of diameter and 150mm of height. There are 24 and 12 samples for concrete grades of M300 and M400, respectively

- Mixing the mixture with a proportion of 1kg INTOC-04 + 8kg of cement + 3kg of water to form waterproofing slurry.

- Spraying water to humidify some concrete samples at the age of 21 day-olds, then covering one side of them with a waterproofing layer of 4mm of thickness (Figure 5);

- After the surface of the oil layer is dry, covering on the waterproofing layer with a protective layer of cement-sand mortar of 1cm (Figure 6);

- Curing the samples with maintaining adequate moisture for 7 days from the moment of casting.

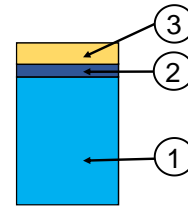


Fig. 5. Sample with waterproofing layer. 1 : concrete sample, 2: waterproofing layer of 4mm, 3 : protective layer of 1cm

A total of 36 seepage tests were conducted according to the Vietnamese norm TCVN 3116: 1993 [21] and BS EN12390 -8-2009 part 8 [22]. Two levels of water pressure of 10daN/cm<sup>2</sup> and 12daN/cm<sup>2</sup> were applied and maintained for 72 hours. After completing the experiments, splitting the sample lengthwise to measure the maximum depth of penetration of water into the concrete. The results are shown in table III and IV:

Results have shown that:

- The waterproofing capacity of the INTOC products for both cases of positive and negative seepage is high. There isn't any seepage flow through the waterproofing layer even at high water pressure of 10 and 12 daN/cm<sup>2</sup>, equivalent to the pressure of water columns of 100 and 120m, respectively.

- In the absence of a waterproofing layer, the concrete of grade M400 has a higher waterproofing capacity than that of M300.



Fig. 6. Samples without waterproofing layer (left), Samples with waterproofing layer (right).

TABLE III. SEEPAGE TESTS ON SAMPLES OF M300 GRADE

N <sup>o</sup>	Experimental pressure (daN/cm <sup>2</sup> )	With/Without waterproofing layer (daN/cm <sup>2</sup> )	Seepage type	Maximum depth of penetration of water into concrete part (mm)	Water permeates through the waterproofing
1	10	Without	Positive	150	NA
2	10	Without	Positive	150	NA
3	10	Without	Positive	150	NA
4	10	Without	Positive	150	NA
5	10	Without	Positive	150	NA
6	10	Without	Positive	150	NA
7	10	With	Positive	0	Not seeped
8	10	With	Positive	0	Not seeped
9	10	With	Positive	0	Not seeped
10	10	With	Positive	0	Not seeped
11	10	With	Positive	0	Not seeped
12	10	With	Positive	0	Not seeped
13	12	With	Positive	0	Not seeped
14	12	With	Positive	0	Not seeped
15	12	With	Positive	0	Not seeped
16	12	With	Positive	0	Not seeped

N <sup>o</sup>	Experimental pressure (daN/cm <sup>2</sup> )	With/Without waterproofing layer (daN/cm <sup>2</sup> )	Seepage type	Maximum depth of penetration of water into concrete part (mm)	Water permeates through the waterproofing
17	12	With	Positive	0	Not seeped
18	12	With	Positive	0	Not seeped
19	10	With	Negative	150	Not seeped
20	10	With	Negative	150	Not seeped
21	10	With	Negative	150	Not seeped
22	10	With	Negative	150	Not seeped
23	10	With	Negative	150	Not seeped
24	10	With	Negative	150	Not seeped

TABLE IV. SEEPAGE TESTS ON SAMPLES OF M400 GRADE

N <sup>o</sup>	Experimental pressure (daN/cm <sup>2</sup> )	With/Without waterproofing layer (daN/cm <sup>2</sup> )	Seepage type	Maximum depth of penetration of water into concrete part (mm)	Water permeates through the waterproofing
1	12	Without	Positive	85	
2	12	Without	Positive	87	
3	12	Without	Positive	92	
4	12	Without	Positive	84	
5	12	Without	Positive	90	
6	12	Without	Positive	88	
7	12	With	Positive	0	Not seeped
8	12	With	Positive	0	Not seeped
9	12	With	Positive	0	Not seeped
10	12	With	Positive	0	Not seeped
11	12	With	Positive	0	Not seeped
12	12	With	Positive	0	Not seeped

### C. Test of adhesion of reinforcement coated with waterproofing slurry and concrete (Pull-out test)

As mentioned in Section 2.3, when INTOC slurry is watered on the surface of the lean concrete layer, the surface of the steel-bar reinforcement of the basement floor can be attached by the slurry. It is necessary to ensure that the adhesion between the steel bars and the concrete floor is not affected by the waterproofing slurry. Therefore, the test of adhesion is essential.

The testing process consists of the following steps:

- Preparing 6 cylindrical molds with a diameter of 150mm and a height of 300mm and 6 rebars with a diameter of 20 mm, yield strength  $f_y = 280$  Mpa (Figure 7).
- Mixing a mixture of waterproofing slurry with a proportion of 1 kg of INTOC + 1 kg of water + 1 kg of cement. Dipping 3 steel bars into the waterproofing slurry oil solution so that the slurry adheres to the surface of them.
- Putting all of 6 steel bars including 3 slurry-coated bars into the molds 200mm, which is 20 times the diameter. Pouring M300 concrete into the molds, curing the samples under standard conditions.
- Using concrete samples with a steel bar plugged at 28 days old for testing.

The experimental results are shown in table 4. For all the tests, the tension force applied to the steel bars continuously increased to the value of ultimate strength. The steel part outside the concrete sample is plastically deformed, then broken. There was no slip between the concrete and the reinforcement. The concrete samples are not vandalized. The maximum pulling force shown in the table corresponds to the ultimate tensile strength of the reinforcement.

Thus, the experimental results have shown that the reinforcement with a glue layer attached on the surface didn't slip off the concrete if the anchor length was greater or equal than 10 times of steel bar diameter.

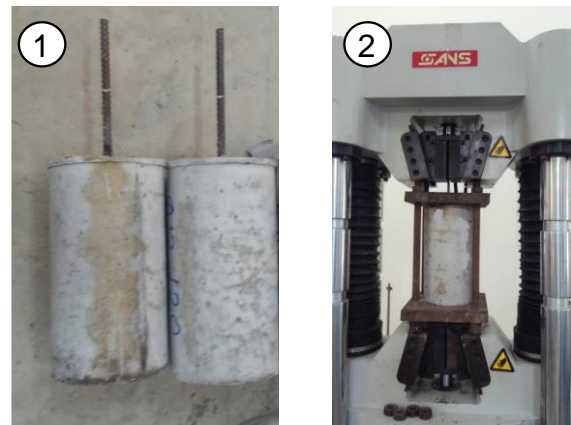


Fig. 7. Adhesion test. 1: concrete samples with reinforcement bars, 2: a sample in test operation

TABLE V. RESULTS OF ADHESION TESTS

N <sup>o</sup>	Reinforcement bars with or without waterproofing on surface	Pullout Strength	
		Strength (kN)	Average (kN)
1	without	86,74	86,25
2	without	85,87	
3	without	86,13	
4	With	86,21	86,12
5	With	85,96	
6	With	86,18	

#### IV. CONCLUSIONS

According to the survey of several buildings and some initial experimental tests, the results have shown that INTOC-04 products have good waterproofing properties, can be used for waterproof as well as reverse waterproofing for basement concrete, with longevity high.

The waterproofing layer on the surface concrete of grade M300 and M400 can prevent water seepage in both positive-side and negative cases at a pressure up to 10 and 12 daN/cm<sup>2</sup> which are equivalent to water column force of 100 to 120 m high, respectively.

The INTOC slurry adhered to the surface of reinforcement does not affect the adhesion between reinforcement and concrete if the anchor length isn't less than 10 times of bar diameter. This is an important property, creating favorable conditions for the waterproofing work for the basement.

The waterproofing has a small adhesion force with water molecules, so water clusters into drops on the surface of the waterproofing layer. Surface tension gives them their near-spherical shape. Even when there is a large crack in the waterproofing sample, the water didn't penetrate through the crack. Therefore, the waterproofing material is hydrophobic, preventing the flow of capillary permeability, reducing the permeability and wetness.

Experiments in this research have just stopped at the level of testing of waterproofing efficiency of products without intensive research on their mechanisms, the absorption of products into the concrete material and the microstructure of waterproofing material and concrete after absorbing waterproofing slurry.

#### REFERENCES

- [1] Duy Hung Nguyen, Hoai Long Le, Minh Tam Nguyen, "Assesing the current statuts of waterpooft in civil construction," Vietnam construction review vol. 12, pp. 26-31, 2016.
- [2] J. Warde, "Really doing something about that damp or wet basement". Available at: <http://www.nytimes.com/1991/08/10/news/really-doing-something-about-that-damp-or-wet-basement.html>, last accessed 2016/07/2016.
- [3] Bob Vila, "Basement renovation - how to waterproof basement - Bob Vila," eps.3402, 2012. Available at: [https://www.youtube.com/watch?v=XR\\_GsF4erpQ](https://www.youtube.com/watch?v=XR_GsF4erpQ), last accessed 2016/7/29
- [4] Bob Vila, "Basement Waterproofing," Bob Vila Radio S. Monzon, ed. Bob Vila, 2013. Available at: <https://www.bobvila.com/articles/bob-vila-radio-asegment-waterproofing>, last accessed 2017/8/11.
- [5] J. Carmody, and B. Anderson, "Moisture in basements: causes and solutions". University of Minnesota extension, 2006. Available at: <http://www.extension.umn.edu/environment/housing-technology/moisture-management/moisture-in-basements-causes-and-solutions/#overview>, last accessed 2016/7/29.
- [6] E. J. Garboczi, "Permeability, Diffusivity and Microstructural Parameters: A critical review," Cement and Concrete Research, vol 20, pp. 591-601, 1990.
- [7] T. C. Powers, "Structures and Physical Properties of hardened portland cement pastes," Journal of the American Ceramic society vol. 41, pp. 1-6, 1958.
- [8] T. C. Powers, L. E. Copeland, H. M. Mann, "Capillary continuity or discontinuity in cement paste," Journal of the PCA research and development laboratories, Bulletin 1 (2), pp. 38-48, 1959.
- [9] J. P. Ollivier, M. Massat, L. Parrott, Transport characteristics, CHAP 4, Performance Criteria for Concrete Durabiliy. JK a.H HK Hilsdorf, (éd.), 1995.
- [10] Khan, MI: Permeation of high performance concrete. Journal of Materials in Civil Engineering, 15 (1), 84-92 (2003).
- [11] M.I. Khan, C.J. Lynsdale, "Strength, permeability, and carbonation of high performance concrete," Cement and Concrete Research, vol. 32 (1), pp. 123-131, 2001.
- [12] B. Gérard, D. Breysse, A. Ammouche, O. Houdusse, O. Didry, "Cracking and permeability of concrete under tension," Materials and Structures, vol. 29 (187), pp. 141-151, 1996.
- [13] P. Mivelaz, Etanchéité des structures en béton armé, fuites au travers d'un élément fissuré. Thèse, Ecole Polytechnique Fédérale de Lausanne, Lausanne, SUISSE, 1996.
- [14] I. Ujike, S. Nagataki, R. Sato, K. Ishikawa, "Influence of internal carcking formed around deformed tension bar on air permeability of concrete," Transactions of the Japan Concrete Institute, vol. 12, pp. 207-214, 1990.
- [15] LION Maxime, SANAHUJA Julien, "Perméabilite a l'eau des betons : développement d'une méthode d'essai alternative par séchage," Conférence Internationale Francophone NoMaD 2018 Liège Université. Liège, Belgique, 7-8 Novembre 2018.
- [16] K. Wang, D. C. Jansen, S. P. Shah, A. F. Karr, "Permeability study of cracked concrete," Cement and Concrete Research, vol. 27 (3), pp. 381-393, 1997.
- [17] Nguyen Tan Van, Letter of certificate, 2016. Available at: <http://www.chongthamintoc.com.vn/khach-hang-noi-ve-chung-toi>, last accessed 2017/8/10.
- [18] Nguyen Truong Luu, Letter of recommendation, 2016. Available at: <http://www.chongthamintoc.com.vn/khach-hang-noi-ve-chung-toi>, last accessed 2017/8/10.
- [19] Dang Thanh Son, Letter of certificate, 2017. Available at: <http://www.chongthamintoc.com.vn/khach-hang-noi-ve-chung-toi>, last accessed 2017/8/10.
- [20] INTOC 2020, Product information and Execution instruction of INTOC-04.
- [21] Ministry of sciences and Technology, TCVN 3116-1993 - Heavyweight concrete -Method for determination of watertightes, 1993.

# Influence of Heating Temperature in Thermal Oxidation to Prepare Titanium Oxide /Aluminum-doped Zinc Oxide Films for Multi-functional-energy-saving Glass

Shang-Chou Chang<sup>1,2</sup>

<sup>1</sup>Department of Electrical Engineering,

<sup>2</sup>Green Energy Technology Research Center,

Kun Shan University

Tainan City, Taiwan

jchang@mail.ksu.edu.tw

Tsung-Han Li

Department of Electrical Engineering,

Kun Shan University,

Tainan City, Taiwan

superjohn0722@yahoo.com.tw

Huang-Tian Chan

Green Energy Technology Research Center

Kun Shan University,

Tainan City, Taiwan

a106000057@g.ksu.edu.tw

**Abstract**— This study reports the feasibility of thermal oxidation to prepare titanium oxide/aluminum-doped zinc oxide (TiO<sub>x</sub>/AZO) films on glass for multi-functional-energy-saving glass: self-cleaning and low emissivity. The titanium/aluminum-doped zinc oxide films were deposited on glass substrates first. Thermal oxidation of on titanium was then carried in a furnace with feeding oxygen gas flow. The heating temperature was set at 300°C, 400°C and 500°C. Structural, optical and hydrophilic properties of the TiO<sub>x</sub>/AZO samples were measured. Results indicate heating temperature in thermal oxidation treatment indeed influence the properties of TiO<sub>x</sub>/AZO significantly especially for 500°C. Particulate agglomeration on surface of TiO<sub>x</sub>/AZO samples increases with heating temperature seen from scanning electron microscope. Zinc substitution with aluminum in zinc oxide is observed from zinc oxide (002) X-ray diffraction peak shifting to high angle when heating temperature increases. The average visible transmittance of TiO<sub>x</sub>/AZO samples increases from 41 % to 81 % as heating temperature raising from 300 to 500° C. The contact angle of TiO<sub>x</sub>/AZO samples after ultraviolet radiation is 7.17° , 8.16° , and 31.26° respect heating temperature: 300, 400 and 500°C. High heating temperature like 500°C makes titanium reacts with oxygen and zinc replace with aluminum more complete in TiO<sub>x</sub>/AZO films. This makes visible transmittance high and emissivity (corresponding to the value of electrical resistivity) low appropriate for low emissivity glass. The good hydrophilic property also observed from low contact angle of TiO<sub>x</sub>/AZO samples reveals the prepared TiO<sub>x</sub>/AZO samples can be also acted as self-cleaning glass.

**Keywords**— Thermal oxidation, Titanium oxide, Aluminum-doped Zinc Oxide

## I. INTRODUCTION

Energy-saving glass is the most widely used in green buildings. Energy-saving glass includes low-emissivity (low-e) glass and self-cleaning glass. Currently, Titanium dioxide (TiO<sub>2</sub>) is the most widely used material in self-cleaning application [1,2]. Titanium (Ti) can be thermally oxidized to form titanium oxide [3]. The photocatalytic properties of TiO<sub>2</sub> can have anti-kill bacteria function (such as coronavirus, H1N1) [4].

H. R. An et al. reported that the photocatalytic property of TiO<sub>2</sub> films can be improved after H plasma treatment, due to the surface modification [5]. Other studies report that doping SnO<sub>2</sub> in TiO<sub>2</sub> films could also improve hydrophilic properties [6]. The hydrophilicity of titanium dioxide is related to the surface microstructure, oxygen vacancies and Ti<sup>3+</sup> [7]. Some work reported [8, 9] that heat process could remove the oxygen in the TiO<sub>2</sub> film. The oxygen vacancies will near hydrophilic properties of the TiO<sub>2</sub> film increase.

On the other hand, low-e glass has high visible light transmission, and has the high infrared light reflectivity property. The emissivity is an important index of low-e glass. The lower low-e glass of emissivity is, the better infrared reflection is for low-e glass. Previous annealing studies on AZO show that emissivity of materials decreases when lowering their electrical resistivity [10].

The TiO<sub>x</sub>/AZO film will be applied to energy-saving glass in this study. Titanium (Ti) films were deposited on glass substrates first by heating Ti films in oxygen atmosphere. The Ti films were thermally oxidized into TiO<sub>x</sub> sample. The temperature of thermal oxidation effect the structure and hydrophilic of TiO<sub>x</sub>. The AZO film layer provides low emissivity property. The AZO film layer will use vacuum-annealing to achieve good emissivity.

## II. EXPERIMENTS

The AZO films were prepared on glass substrates in sputtering system. The AZO films have thickness in 500 nm. The AZO layer was sputtered using an AZO target (ZnO:Al<sub>2</sub>O<sub>3</sub> = 98:2 wt.%) at DC power of 2 kW, with Ar flow of 440 sccm and working pressure of 3×10<sup>-3</sup> Torr. After that, the AZO films were post-annealed in vacuum. The AZO films were vacuum-annealed at 400°C for 1 hour, and working pressure of 3×10<sup>-5</sup> Torr. The carrier concentration, mobility, and electrical resistivity of AZO films were obtained by the Hall measurement (Hall effect measurement, Ecopia HMS-3000, Ecopia, gyeonggi-do, South Korea). Emissivity calculated from the Hagen-Rubens relation by electrical resistivity [11].

Ti films were deposited on as-deposited AZO films by electron beam deposition system to fabricate Ti/AZO films.

The electron beam deposition system using the parameters: output power: 8 kW, turntable speed: 6 rpm, thickness of films: 100 nm. After deposition, the Ti/AZO films were thermally oxidized in rapid thermal annealing system, with feeding of 100 sccm oxygen, and heating rate at 25°C. The Ti/AZO sample was thermally oxidized with 300°C, 400°C, and 500°C by 10 minutes respectively into TiO<sub>x</sub>/AZO sample. The TiO<sub>x</sub>/AZO films were different temperature at 300°C, 400°C, and 500°C named as sample S<sub>1</sub>, S<sub>2</sub> and S<sub>3</sub> respectively.

The crystalline structure of TiO<sub>x</sub>/AZO films was analyzed by X-ray diffractometer (XRD; Rigaku D/Max2500, Rigaku, Tokyo, Japan). The surface morphology of TiO<sub>x</sub>/AZO films was observed using Scanning Electron Microscope (SEM; Hitachi SU8000, Hitachi, Tokyo, Japan). The surface roughness of TiO<sub>x</sub>/AZO films was measured using Scanning Probe Microscopes (SPM; Bruker ICON3-SYS, Bruker, Siegsdorf, Germany). The optical transmittance of TiO<sub>x</sub>/AZO films was measured using a UV/VIS/NIR spectrophotometer (PerkinElmer LAMBDA 750, PerkinElmer, Waltham, U.S.A.) in the wavelength range of 380~780 nm. The hydrophilicity of TiO<sub>x</sub>/AZO films were measured using a contact angle meter (First Ten Angstroms FTA 1000B, First Ten Angstroms, Portsmouth, U.S.A.).

### III. RESULTS AND DISCUSSION

Figure 1 shows the XRD spectra of the TiO<sub>x</sub>/AZO films. The TiO<sub>x</sub>/AZO films sample had the highest diffraction peak at (002) and (103), which corresponds to the zinc oxide crystal plane. The peaks corresponding to (002) and (103) ZnO crystal plane were observed in Table 1. But any TiO<sub>x</sub>/AZO films do not have the signal of TiO<sub>2</sub>. It may be that AZO films signal cover was anatase peak signal. These studies result like to R.C. Suci et al [15]. According to R.C. Suci et al. reported that the TiO<sub>2</sub> films annealing temperature at above 500°C observed anatase phase appears; has not anatase peak signal below 500°C. From the Table 1, the peak with respect to (002) is shift to a higher angle with increasing heating temperature. The peak with respect to (002) increases when after different temperature, which may imply that Al ions replace the lattice position of Zn ions. The (002) crystal plane spacing decrease, the possible causes the lattice position of Zn ions occupied by Al ions. The phenomenon will provide free electrons such that the electrical resistivity of the AZO decreases [12]. According to Y. Liu et al. [13], the Al ions are replaced by Zn ions during annealing of Zn/AZO films, the distance between crystal plane is reduced and the electrical properties is increased of films.

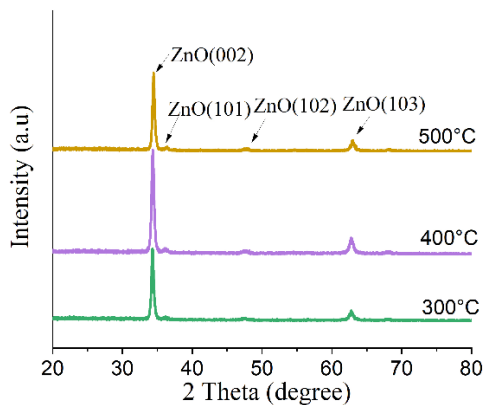


Fig. 1. XRD spectra of the thermal oxidation processed TiO<sub>x</sub>/AZO films at temperatures of (a) 300°C, (b) 400°C, and (c) 500°C.

TABLE I. STRUCTURE AND OPTICAL PROPERTIES OF THE THERMAL OXIDATION PROCESSED WITH DIFFERENT TEMPERATURES OF TiO<sub>x</sub>/AZO FILMS.

Sample	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>
Treatment temperatures	300 °C	400 °C	500 °C
ZnO(002) 2θ (°)	34.31°	34.33°	34.45°
ZnO(103) 2θ (°)	62.77°	62.76°	62.98°
Average roughness (R <sub>a</sub> ) nm	4.04	4.78	7.45
Average optical transmittance (%)	41.47%	69.08%	81.12%

The average roughness (R<sub>a</sub>) of the TiO<sub>x</sub>/AZO films measured by SPM was shown in Table 1. Higher temperature increases average roughness of TiO<sub>x</sub>/AZO films with heat oxidation. The average roughness of the S<sub>2</sub> and S<sub>3</sub> films increases 4.78 nm and 7.45 nm compared with that of the S<sub>1</sub> films, respectively. Observed thermally oxidized TiO<sub>x</sub>/AZO films surface morphology by the SEM. The surface micrograph of the S<sub>1</sub>, S<sub>2</sub>, and S<sub>3</sub> is respectively shown in Figure 2. The surface grain size change with different temperatures. The sample of Figure 2 (b) has a dense grain structure. With increase in processing temperature from 300 to 500°C, the grain structures become agglomeration. About structural change, the surface morphology for S<sub>1</sub>, S<sub>2</sub> and S<sub>3</sub> TiO<sub>x</sub>/AZO films are similar to those for average roughness.

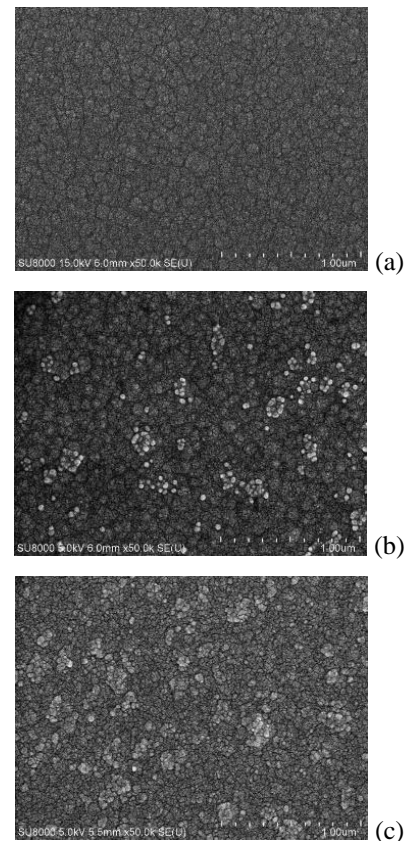


Fig. 2. Scanning electron micrographs of the thermal oxidation processed TiO<sub>x</sub>/AZO films at temperature of (a) 300, (b) 400, and (c) 500°C. With increase in processing temperature from 300 to 500°C, the grain structures become agglomeration.



Figure 3 show the transmission spectra in visible light range of the TiO<sub>x</sub>/AZO films. The average transmittance for the TiO<sub>x</sub>/AZO films is shown in Table 1. The average transmittance in visible light of S<sub>1</sub>, S<sub>2</sub>, and S<sub>3</sub> was 41.47%, 69.08%, and 81.12%, respectively. Higher temperature increases average transmittance of TiO<sub>x</sub>/AZO films with heat oxidation. The heating temperature is up to 400 and 500 °C, making titanium react with oxygen. It may be that Ti metal were transformed to the ceramic material TiO by thermal oxidation [14].

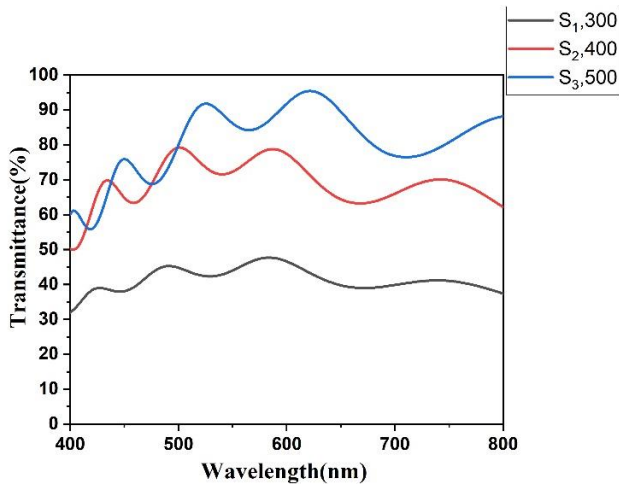


Fig. 3. Optical transmittance spectra of the thermal oxidation processed with different temperatures of TiO<sub>x</sub>/AZO films.

Table 2 shows the electrical properties and calculated emissivity of the different AZO films. The electrical resistivity of vacuum-annealed AZO films reduces  $1.04 \times 10^{-3} \Omega\text{-cm}$  compared to that of as-deposited films. The emissivity experimental data shall be calculated using the following Hagen-Rubens relation [11]. Equation (1) is Hagen-Rubens relation.

$$\varepsilon = 0.0129 \times R - 6.7 \times R^2 \quad (1)$$

where R is electrical resistivity of AZO films. Emissivity computed from the Hagen-Rubens relation, the emissivity of as-deposited and vacuum-annealed AZO films was 0.6, and 0.24, respectively. The emissivity of the AZO films decreases with the decrease of the electrical resistivity. This means that the TiO<sub>x</sub>/AZO films after annealing, the better infrared reflection is for low-e glass.

TABLE II. ELECTRICAL PROPERTIES AND EMISSIVITY OF AZO FILMS THAT ARE PRODUCED USING DIFFERENT METHODS.

Process Conditions	as-deposited	vacuum-annealed
Carrier concentration ( $10^{20}/\text{cm}^3$ )	2.5	4.1
Carrier mobility ( $\text{cm}^2/\text{Vs}$ )	6.39	14.2
Resistivity ( $10^{-3}\Omega\text{-cm}$ )	3.93	1.04
Emissivity	0.60	0.24

The contact angle of S<sub>1</sub>, S<sub>2</sub>, and S<sub>3</sub> is shown in Table 3. Experiment results indicates applied temperature in thermal oxidation of TiO<sub>x</sub>/AZO samples indeed influences the contact angle. At first, the contact angles were 56.90°, 62.83° and 68.00° for the films at S<sub>1</sub>, S<sub>2</sub>, and S<sub>3</sub>, respectively. After UV irradiation for 1 hour, the contact angles reduce to 7.17°, 8.16°, and 31.26°, respectively. When the sample was applied temperature below 400°C in thermal oxidation, the contact angle was lower than 10°. However, when the sample was applied temperature at 500°C, leading to the increase in the contact angle. The photoinduced hydrophilicity of TiO<sub>2</sub>, may be attributed to the hydroxyl radical groups and oxygen vacancies of surface. When TiO<sub>2</sub> is exposed to UV light, electron-hole pairs are generated. The holes can react with water to produce the hydroxyl radical ( $\bullet\text{OH}$ ). Electrons can also change the Ti<sup>4+</sup> to Ti<sup>3+</sup> state; then form oxygen vacancies. Water can then occupy oxygen vacancies; produce to the hydroxyl radical. These hydroxyl radical adsorption of water on the TiO<sub>2</sub> surface [7, 9]. Contact angle was affected upon thermal oxidation of different temperature. These results are similar to those of F. Meng et al [9]. F. Meng et al. reported that the TiO<sub>2</sub> surface has hydrophilic, may be attributed to annealing treatment could form the oxygen vacancy.

TABLE III. CONTACT ANGLE OF THE THERMAL OXIDATION PROCESSED WITH DIFFERENT TEMPERATURES OF TiO<sub>x</sub>/AZO FILMS.

TiO <sub>x</sub> /AZO films irradiation and without UV Radiation			
Sample	Treatment temperatures	No UV	Irradiation UV
S <sub>1</sub>	300 °C	56.90°	7.17°
S <sub>2</sub>	400 °C	62.83°	8.16°
S <sub>3</sub>	500 °C	68.00°	31.26°

#### IV. CONCLUSION

In this study, the TiO<sub>x</sub>/AZO film was subjected to the thermal oxidation processed with different temperatures (300°C, 400°C, and 500°C). The structure transform as temperature increases due to thermal oxidation treatment, which increases average transmittance. The surface morphology of the TiO<sub>x</sub>/AZO film observed by SEM. From the average roughness and figure 2, the process temperature was that the grains become small and dense phenomenon. The annealing decreases the electrical resistivity and lowers the emissivity of TiO<sub>x</sub>/AZO films. The emissivity of the fabricated multi-functional-energy-saving glass with the vacuum-annealed samples is 0.24 which is much lower than that with the as-deposited samples and nearly 2.5 times lower than that with the as-deposited TiO<sub>x</sub>/AZO films. The average transmittance in visible light of TiO<sub>x</sub>/AZO films higher as the applied temperature in thermal oxidation of TiO<sub>x</sub>/AZO films increases. Compared to sample at 300°C, and 400°C, in the environment of 500°C has a best average transmittance in visible light. In addition, 300°C and 400°C has a good hydrophilic, the contact angle was lower than 10°. In this study investigated the feasibility of applying thermal oxidation TiO<sub>x</sub>/AZO films to multi-functional-energy-saving glass.

# ACKNOWLEDGMENT

The authors thank the Ministry of Science and Technology (MOST) and Ministry of Education, Taiwan for financial support (MOST 105-2221-E-168-101), (MOST 106-2221-E-168-021) and (107-N-270-EDU-T-142).

# REFERENCES

- [1] K. Midtdal, B. P. Jelle, "Self-cleaning glazing products: A state-of-the-art review and future research pathways," *Sol. Energy Mater. Sol. Cells*, vol. 109, pp. 126-141, February 2013.
- [2] M. J. Miller, J. Wang, "Multilayer ITO/VO<sub>2</sub>/TiO<sub>2</sub> thin films for control of solar and thermal spectra," *Sol. Energy Mater. Sol. Cells*, vol. 154, pp. 88-93, September 2016.
- [3] A. Trenczek-Zajac, "Influence of etching on structural, optical and photoelectrochemical properties of titanium oxides obtained via thermal oxidation," *Mater Sci Semicond Process*, vol. 83, pp. 159-170, August 2018.
- [4] T. An, H. Zhao, and P. K. Wong, *Advances in Photocatalytic Disinfection*, 1st ed., Germany: Springer, 2017, pp. 7-9.
- [5] H.-R. An, S. Y. Park, H. Kim, C. Y. Lee, S. Choi, S. C. Lee, S. Seo, E. C. Park, Y.-K. Oh, C.-G. Song, J. Won, Y. J. Kim, J. Lee, H. U. Lee, and Y.-C. Lee, "Advanced nanoporous TiO<sub>2</sub> photocatalysts by hydrogen plasma for efficient solar-light photocatalytic application," *Sci. Rep.*, vol. 6, pp. 29683, July 2016.
- [6] M. Farbod, S. Rezaian, "An investigation of super-hydrophilic properties of TiO<sub>2</sub>/SnO<sub>2</sub> nano composite thin films," *Thin Solid Films*, vol. 520, pp. 1954-1958, January 2012.
- [7] S. Banerjee, D. D. Dionysiou, and S. C. Pillai, "Self-cleaning applications of TiO<sub>2</sub> by photo-induced hydrophilicity and photocatalysis," *Appl. Catal. B*, vol. 176-177, pp. 396-428, October 2015.
- [8] Y. Sun, S. Sun, X. Liao, J. Wen, G. Yin, X. Pu, Y. Yao, and Z. Huang, "Effect of heat treatment on surface hydrophilicity-retaining ability of titanium dioxide nanotubes," *Appl. Surf. Sci.*, vol. 440, pp. 440-447, May 2018.
- [9] F. Meng, L. Xiao, and Z. Sun, "Thermo-induced hydrophilicity of nano-TiO<sub>2</sub> thin films prepared by RF magnetron sputtering," *J. Alloys Compd.*, vol. 485, pp. 848-852, October 2009.
- [10] S.-C. Chang, and H.-T. Chan, "Effect of nitrogen flow in hydrogen/nitrogen plasma annealing on aluminum-doped zinc oxide/tin-doped indium oxide bilayer films applied in low emissivity glass," *Crystals*, vol. 9, pp. 6, June 2019.
- [11] E. Hagen, H. Rubens, "Über Beziehungen des Reflexions- und Emissionsvermögens der Metalle zu ihrem elektrischen Leitvermögen," *Ann. Phys.*, vol. 11, pp. 873-901, April 1903.
- [12] H. Tong, Z. Deng, Z. Liu, C. Huang, J. Huang, H. Lan, C. Wang, Y. Cao, "Effects of post-annealing on structural, optical and electrical properties of Al-doped ZnO thin films," *Appl. Surf. Sci.*, vol. 257, pp. 4906-4911, March 2011.
- [13] Y. Liu, S. Zhu, B. Song, "Magnetron sputtering deposition of Zn/AZO multilayer films: Towards the understanding of Zn diffusion in AZO film," *Results Phys.*, vol. 13, pp. 102286, June 2019.
- [14] M. A. Butt, S. A. Fomchenkov, "Thermal Effect on the Optical and Morphological Properties of TiO<sub>2</sub> Thin Films Obtained by Annealing a Ti Metal Layer," *J. Korean Phys. Soc.*, vol. 70, pp. 169-172, January 2017.
- [15] R. C. Suci, E. Indrea, T. D. Silipas, S. Dreve, M. C. Rosu, V. Popescu, G. Popescu and H. I. Nascu, "TiO<sub>2</sub> thin films prepared by sol-gel method," *J. Phys.: Conf. Ser.*, vol. 182, pp. 012080, September 2009.

# Comparative Studies of Different Methods for Short-term Locational Marginal Price Forecasting

Ying-Yi Hong

Department of Electrical Engineering  
Chung Yuan Christian University  
Taoyuan City, 32023, Taiwan  
yyhong@ee.cycu.edu.tw

Rolando Pula

Department of Electrical Engineering  
Chung Yuan Christian University  
Taoyuan City, 32023, Taiwan

**Abstract**—This paper presents studies on short-term locational marginal price (LMP) forecasting using six common methods, namely, Persistence, Autoregression Integrated Moving Average (ARIMA), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Convolutional Neural Network-LSTM (CNN-LSTM), and CNN-Fully-Connected Network (CNN-FCN). These six methods were used to study one location in the 2008–2018 datasets of PJM (Pennsylvania, New Jersey, Maryland) power market. Simulation results show that performance of each method is hyperparameter-dependent. Also, the deep learning-based LSTM, CNN-LSTM and CNN-FCN outperform the supervised learning-based RNN. The traditional ARIMA and persistence usually attained the worst accuracy in terms of coefficient of determination and normalized mean squared error.

**Keywords**—Electricity Price Forecasting, Persistence, ARIMA, Supervised Learning, Deep Learning.

## I. INTRODUCTION

Electric Price Forecasting (EPF) has been a main research in the power market for the past years because of increasing demand on electricity around the world. In fact, the U.S Energy Information Administration in 2017 projected that world energy use will increase by 28% in 2040. Additionally, smart grid development, which allows the two-way flows of valuable information from and into the electricity consumers and producers, is also a crucial factor for increase of studies in EPF [1]. In the modern smart grids, consumers can access the real-time pricing of electricity, which is important in consumption schedule in an hourly basis; likewise, producers can adjust and re-allocate generations and supplies of electricity for increasing their profits [2].

Locational Marginal Pricing (LMP) as defined in [3] is the additional MW load supplied at each bus location at a certain price. LMP is important in determining how much electricity the utility needs to acquire in order to become profitable while having a stable system [3]. At an electrical connection point, LMP is the result caused by balance between buyers and sellers in a bidding system for an amount of power [4]. Forecasting the LMP is not easy due to the factors affecting its variations [5]. Some of these major factors were mentioned in [3]. Currently, several methods have been developed for forecasting LMP including statistical approach (SA), machine learning (ML), and deep learning (DL) neural networks. Example of application of SA for EPF can be found in [6]. In [5] the authors compared some ML and DL models' performances for predicting LMP with hourly intervals while [4] compared Auto-Regression Integrated Moving Average

(ARIMA), rolling average, and Long Short-Term Memory (LSTM) using 1-year data of LMP with 5 minutes' interval. In the review conducted by [7], it can be seen that the neural network (NN) based model used for EPF is more dominant than SA by volume of citations and publications. A study showed that DL outperformed other existing SA methods in the day-ahead EPF [8]. However, to the best knowledge and effort of the authors, no study was conducted yet, if it is true in large dataset for 1-hour ahead LMP forecasting. This paper will compare the performance obtained by two SA models and the common ML/DL neural networks used for time-series forecasting.

The rest of the paper is organized as follows. Section II presents a detailed description about LMP data preprocessing. Section III presents implementation of various methods, which are Persistence, ARIMA, Recurrent Neural Network (RNN), LSTM, Convolutional Neural Network-LSTM (CNN-LSTM), and CNN-Fully-Connected Network (CNN-FCN). Section IV presents the results of simulations using a 10-year dataset from the PJM power market. Section V draws conclusions.

## II. DATA PRE-PROCESSING

A systematic way of conducting this study is shown in Fig. 1. First, data was checked and analyzed before putting into the testing models. The data preprocessing consists of four steps: data grouping as well as evaluations of coefficient of variation, normality and stationary. After the data preprocessing, the statistical methods, ML and DL neural networks will be tuned or optimized to obtain results and draw conclusions.

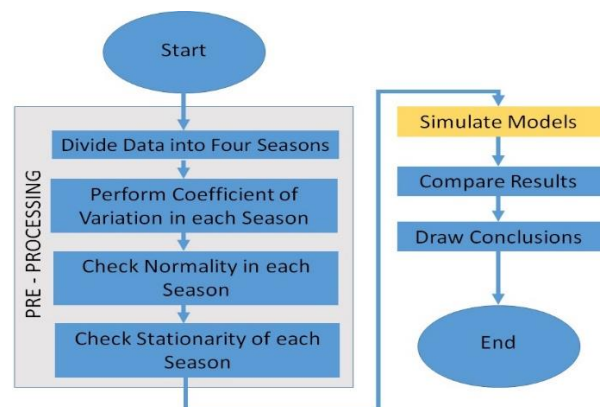


Fig. 1. Overview of the entire study.

The LMPs of (Dayton transmission zone) in the PJM power market were divided into four subsets according to seasons, following the length of season based from the American standard as follows: June – August (summer), September – November (fall), December – February (winter), and March – May (spring). After the LMPs are divided into four subsets, statistical test is used to examine the coefficient of variation (CV) for each season. CV is important to know which season should be used as a reference for optimizing the models. CV is the ratio of the Standard Deviation ( $\sigma$ ) to the mean ( $\mu$ ). The greater the value of the CV, the more the variability of the data, indicating that it is harder to generalize the characteristics of the corresponding time-series data. As can be seen from Table I, the winter dataset is a good candidate as a reference for tuning the studied models since it has the highest value of CV.

TABLE I COEFFICIENTS OF VARIATION

Season	$\sigma$	$\mu$	$\sigma/\mu$
Winter	45.54	43.97	1.036
Fall	19.59	34.85	0.562
Spring	27.50	38.39	0.716
Summer	31.62	40.86	0.774

Two other tests are performed to examine the characteristics of time-series data. One is the D'Agostino's K-squared test, which is a goodness-of-fit measure of departure from normality in terms of kurtosis and skewness [9]. Another is the Augmented Dickey-Fuller (ADF) Test, which examines null hypothesis or determines if the dataset needs to undergo transformation before inputting into the test models [12]. Equations (1) and (2) stand for skewness and kurtosis, respectively.

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}} \quad (1)$$

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3 \quad (2)$$

where  $x_i$  denotes a sample of  $n$  observations;  $\bar{x}$  is the mean of the samples. For  $g_1$ , the following rules will be applied: if  $-0.5 < g_1 < 0.5$ , the data are fairly symmetrical; if  $-1.0 < g_1 < -0.5$  or  $0.5 < g_1 < 1.0$ , the data are moderately skewed, and if  $g_1 < -1$  or  $g_1 > 1$  the data are highly skewed [11].  $g_2$  is the heaviness of the tails of a normal distribution; thus, if  $g_2 > 0$  then distribution has a heavier tail, and if  $g_2 < 0$  then the distribution has a light tail [11].

Aside from  $g_1$  and  $g_2$ , a hypothesis is set for p-value as follows: if p-value  $\leq 0.5$ , reject the null hypothesis; in other words, the distribution is not normal. If p-value  $> 0.5$ , fail to reject null hypothesis, implying the distribution is normal. Table 2 summarizes the values of  $g_1$ ,  $g_2$  and p-value for four seasons.

TABLE II. RESULTS OF D'AGOSTINO'S K<sup>2</sup> TEST

Season	$g_1$	$g_2$	p-value
Winter	435.3152	14.58776	0
Fall	18.03845	3.345611	0
Spring	138.0417	7.632894	0
Summer	81.07001	5.445569	0

As can be seen in Table II, all  $g_1$ 's are greater than one, which means that all datasets in four seasons are highly skewed. Besides, all values in four seasons for  $g_2$  are greater than zero, which implies that the datasets have heavier tails. Because all p-values are nearly zero, all four seasons are not normally distributed.

The next step is to perform ADF, which determines the trends over time. If the trends are fluctuating over time, then transformation is needed to remove any trend before performing any other operation. Data transformation is helpful in removing any trend in the data. This is helpful to have good results when machine learning models are explored. Equations (3), (4) and (5) summarizes the concept behind the Dickey – Fuller (DF) and ADF tests. Given the data  $y_t$ , wherein

$$y_t = \alpha + \beta t + \Phi y_{t-1} + e_t \quad (3)$$

If  $\Phi = 0$  in the given data, (3) can be written as below:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta t + \gamma y_{t-1} + e_t \quad (4)$$

Equation (4) is a linear regression of  $\Delta y_t$  against  $t$  and  $y_{t-1}$  and is used to test whether  $\gamma$  is different from 0. If  $\gamma = 0$ , then this dataset is not stationary (random walk process); if not and  $-1 < 1 + \gamma < 1$ , then the data is stationary. For a higher-order autoregressive process, DF is extended to ADF by including  $\Delta y_{t-p}$  by leaving the test the same as  $\gamma = 0$ , which turns into (5).

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \dots \quad (5)$$

In this paper, the null hypothesis for ADF of all four seasons is that time series is not stationary. The alternative hypothesis is time series of all four seasons is stationary. Table III shows the results of ADF for all four seasons. The interpretation can be given by examining the p-value as follows. If p-value  $> 0.05$ , then it fails to reject the null hypothesis (the data is non-stationary). If p-value  $\leq 0.05$ , then reject the null hypothesis (the data is stationary).

TABLE III. ADF TESTS AND THEIR P-VALUES

Season	ADF statistics	p-value
Winter	-14.1801	0.0007
Fall	-8.06536	0.0001
Spring	-11.0075	0.0143
Summer	-11.0159	0

According to Table III, all p-values are less than 0.05 which reject the null hypothesis. Therefore, the data transformation is not required for all seasons.

### III. IMPLEMENTATION OF VARIOUS METHODS

This section will discuss all the methods/models used to perform the EPF. The baseline method used is the persistence. A baseline model is important in knowing how well the other methods perform. All tuned parameters and hyper-parameters used in each method will be discussed, too. This section provides a concise explanation of the functionality of those parameters and hyper-parameters. Input data in all methods were normalized within [0, 1]. In addition, 80% of the data in each season was used for training while the remaining 20% was used for testing.

### A. Persistence

Persistence method was used as a baseline because it is fast, simple, and can be repeated easily. Following the representation of [12], the LMP at time  $t + 1$  will be equal to LMP at time  $t$ . Persistence method is also called naïve predictor.

$$LMP_{t+1} = LMP_t \quad (6)$$

Similarly, (6) can also be extended to predict the future outcome at  $(t + 1)$  using the previous LMP at  $(t - 1)$ .

### B. ARIMA

ARIMA is one of the frequently used statistical methods for forecasting time series data. Its acronym stands for three operations involved [13]. The first one is the Autoregression (AR), which uses the relationship of dependency between number of some lagged observations and the actual observation [13]. The second one is the Integration (I), which applies differencing to the raw observations to make the time series stationary [13]. Lastly, the Moving Average (MA) is applied to the lagged observations and residual error [13]. Equation (7) explains the mathematics behind ARIMA. Given a time series data  $X_t$ , an  $ARIMA(p, d, q)$  can be expressed as follows.

$$(1 - \sum_{i=1}^p \alpha_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^q \phi_i L^i) \varepsilon_t \quad (7)$$

where  $L$  is the lag operator;  $\alpha_i$  is the parameter of autoregressive part;  $\phi_i$  is the parameter of the MA part and  $\varepsilon_t$  is the error term. Three parameters are needed to be set in ARIMA. They are  $(p, d, q)$ ;  $p$  is lag observation which is also called the lag order;  $d$  is the number of times the raw observations undergo differencing which is also called the degree of differencing, and  $q$  is the window size of MA which is also called the order of MA.

ARIMA model was tested on the 20% of the data in each season with the set parameters of  $(p, d, q) = (5, 1, 0)$ . Results of the experiments can be found in the next section.

### C. RNN

RNN is one of the most frequently used neural network for time series forecasting because of its capability in forming directed cycle. RNN can retain its previous state to the next using the output as input to perform the next step. Fig. 2 shows the structure of a simple RNN.

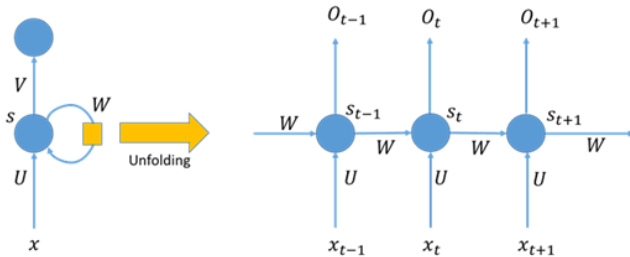


Fig. 2. RNN basic structure.

As shown in Fig. 2,  $x_t$  is the input at time step  $t$ ;  $s_t$  is the hidden state at time  $t$  and  $o_t$  is the output state at time  $t$ . In order to obtain the output state  $o_t$ , the previous state ( $s_{t-1}$ ) of the output  $o_{t-1}$  was used as an input.

The configuration of RNN used in this paper was based on that in [14]; specifically, the numbers of input neurons, hidden layers, neurons in the hidden layer, output neurons are 3, 1, 2 and 1, respectively. Optimizer used is Adaptive Moment Estimation (Adam) and loss function used is mean squared error. The maximum epoch is 100 and batch size is 24.

### D. LSTM

The LSTM is also one of the most popular DL networks used in time-series forecasting. Moreover, it is an enhancement of RNN [15]. Fig. 3 is the structure of LSTM. The operation can be expressed in (8)~(13) [16]:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (8)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (9)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (10)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (11)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (12)$$

$$h_t = o_t \times \tanh(C_t) \quad (13)$$

where  $x_t$  is the input to the network;  $h_t$  is the hidden layer output.  $\sigma$  signifies the activation function.  $C_t$  denotes the state of a cell and  $\tilde{C}_t$  means the candidate state value. The main difference between LSTM and RNN is that LSTM has three gates. The input gate decides whether the information will be retained or not; the forget gate is responsible for determining if the information will be ignored or not. Recording of the processing state will be in the cell, and the output of LSTM values will be passed through the output gate. Through this process, the LSTM has the ability in learning long-term dependencies in time-series data. Specifically,  $W_f$ ,  $W_i$  and  $W_o$  denote the weights of forget gate, input gate and output gate, respectively, while  $W_C$  is the weight of the cell.  $b_f$ ,  $b_i$ ,  $b_o$ , and  $b_C$  stand for the biases of the gates and cell.  $f_t$ ,  $i_t$  and  $o_t$  represent the forget gate, input gate and output gate, respectively.

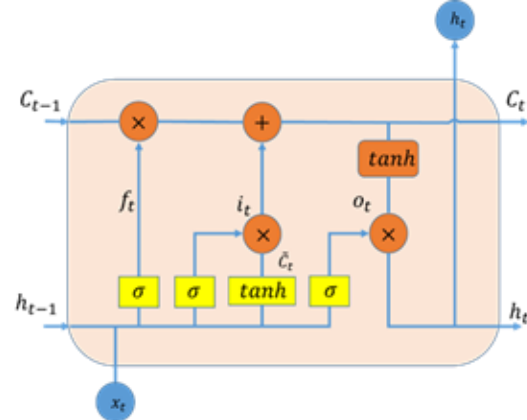


Fig. 3. LSTM structure.

After fine tuning the LSTM network used, the optimal numbers of input cells, LSTM layer, cells inside the LSTM layer and output neurons are 30, 1, 200 and 1, respectively.



The dropout value is 0.2. The optimizer Adam was used and the loss function was the mean squared error. The batch size and the maximum of epochs were set to 24 and 100, respectively.

#### E. CNN-FC

The CNN is mainly used for feature extraction and data pre-processing using the grid topology [17]. The CNN uses a special operation called convolution, which is applied to functions of real valued arguments [18]. This can be formulated in (14).

$$feature = f(x, w) = x * w \quad (14)$$

where  $x$  is input and  $w$  is the weighing function – also called “kernel”. The asterisk is the convolution operation, and  $feature$  is the “feature map”. The kernel or filter as defined in [19] is an array of weights which change with the learning algorithm through iterations.

Another operation involving in the CNN is the pooling operation. This operation is usually performed in the 2nd stage of CNN. Pooling operation is used to smoothen and modify the feature map [19]. There are usually four kinds of pooling operation/function used in literatures [18]. These are the minimum pooling, maximum pooling, weighted average pooling, and average pooling. The maximum pooling attains the maximum value inside a defined window size. The average pooling is designed to obtain the average value of a defined window size. The minimum pooling gains the minimum value in the defined window and the weighted average pooling results in the weighted average of the values inside a defined window size [18].

Stride is a scheme to conduct the convolution. This operation defines how many steps the kernel operation should move to perform the convolution each time.

The CNN is usually cascaded by a fully-connected network (FCN) where the regression operation is performed. As shown in Fig. 4, the structure of fined-tuned 1D CNN-FCN was used in this paper. The input shape was  $24 \times 1$  with 5 filters of  $3 \times 1$  in length. The number of convolution used was 128 and the average pooling operation was used with a window size of 2. The stride of 1 was used. The dropout value of 0.3 was used after the flatten layers to avoid over-fitting. The FCN layer consists of 1 dense layer with 100 neurons. The activation function used in the FCN side is Rectified Linear Unit (ReLU). This CNN-FCN was compiled using mean squared error as the loss function optimized by Adam.

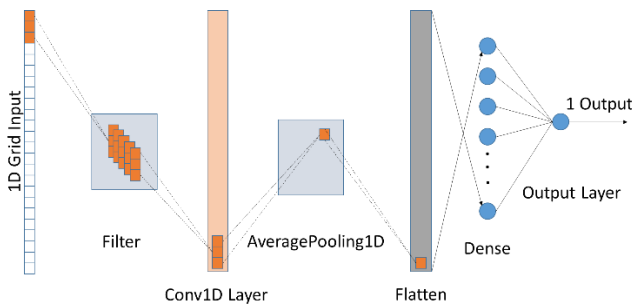


Fig. 4. Architecture of 1D CNN-FCN.

#### F. CNN-LSTM

CNN-LSTM is the integration of CNN with LSTM. The CNN part is for feature extraction while the LSTM is for forecasting part [5]. As shown in Fig. 5, the structure of CNN-LSTM is based on that presented in [5].

The input shape of the architecture in Fig. 5 is  $24 \times 1$  with kernel size of  $1 \times 1$  and 32 convolutions in the first convolution layer. ReLU was the activation function used in the 1<sup>st</sup> convolution layer. The maximum pooling was applied after the 1<sup>st</sup> convolution with a window size of  $2 \times 1$ . Another convolution was performed after the first pooling layer with 32 convolutions of kernel size  $1 \times 1$ . Batch normalization was applied after the 2<sup>nd</sup> convolution layer. Again, ReLU was used as the activation function for the 2<sup>nd</sup> convolution layer. The maximum pooling with a window size of  $2 \times 1$  was performed after the 2<sup>nd</sup> convolution. The output of the 2<sup>nd</sup> maximum pooling was flattened and was connected to the LSTM with 32 cells. ReLU was the activation function used in the LSTM. The LSTM output was then connected to a single neuron as a single output.

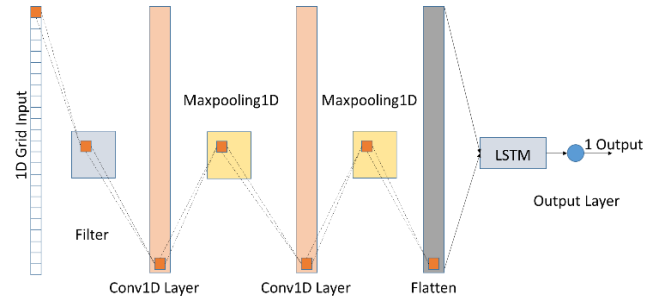


Fig. 5. Architecture of CNN-LSTM used in [5] followed in this paper.

### IV. RESULTS AND DISCUSSION

#### A. Data Description

The data used in this paper were downloaded from Pennsylvania, Jersey, Maryland (PJM) power pool. It was a real-time hourly LMP dataset for Dayton Transmission Zone from January 1, 2008 to December 31, 2018. The plot of the whole data is shown in Fig. 6.

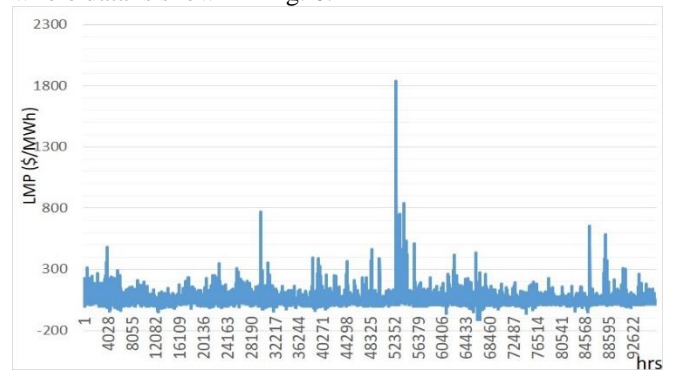


Fig. 6. Ten years' LMPs.

#### B. Evaluation Metrics and Results

This section discusses the results obtained from the different methods described in Sec. III. Four performance metrics were used to evaluate all methods. These are coefficient of determination ( $R^2$ ), Normalized Mean Squared Error (NMSE), Root Mean Squared Error (RMSE), and Mean

Absolute Percentage Error (MAPE) which are formulated in (15)~(18) [19, 20, 21].

$$R^2 = 1 - \frac{\sum(\hat{y}-y)^2}{\sum(y-\bar{y})^2} \quad (15)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \quad (16)$$

$$NMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}}{\bar{y}} \quad (17)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (18)$$

where  $\hat{y}$  is the predicted value,  $\bar{y}$  is the mean of the predicted values, and  $y$  is for actual value.  $R^2$  is used to measure the linear correlation between the actual and predicted values. When  $R^2$  is close to 1, it indicates that predicted values are close to actual values. MAPE, NMSE and RMSE quantify the measured errors between the actual and predicted values. Tables IV~VII show the results of Persistence, ARIMA, RNN, LSTM, CNN-LSTM, and CNN-FCN for all seasons.

TABLE IV. PERFORMANCE OF VARIOUS METHODS FOR WINTER DATASET

Model	RMSE (\$/MWh)	NMSE	MAPE (%)	$R^2$
PERSISTENCE	19.845	0.5336	12.658	0.5907
ARIMA	19.245	0.5173	13.446	0.8533
RNN	18.775	0.5191	<b>12.187</b>	0.8553
LSTM	18.026	0.5300	14.481	0.8724
CNN-LSTM	18.005	0.4924	14.512	0.8718
CNN-FCN	<b>17.510</b>	<b>0.4662</b>	15.285	<b>0.8750</b>

TABLE V. PERFORMANCE OF VARIOUS METHODS FOR SPRING DATASET

Model	RMSE (\$/MWh)	NMSE	MAPE (%)	$R^2$
PERSISTENCE	14.686	0.4738	<b>19.821</b>	0.6640
ARIMA	14.075	0.454	21.935	0.6555
RNN	13.489	0.4353	20.391	0.6691
LSTM	<b>12.968</b>	<b>0.3819</b>	26.358	<b>0.7101</b>
CNN-LSTM	13.381	0.4112	25.064	0.6829
CNN-FCN	13.722	0.4466	21.252	0.6433

TABLE VI. PERFORMANCE OF VARIOUS METHODS FOR SUMMER DATASET

Model	RMSE (\$/MWh)	NMSE	MAPE (%)	$R^2$
PERSISTENCE	11.937	0.4045	14.565	0.6879
ARIMA	11.319	0.3835	15.118	0.7020
RNN	10.972	0.3633	15.426	0.7036
LSTM	<b>9.8399</b>	0.3376	11.847	0.7615
CNN-LSTM	10.364	<b>0.3287</b>	15.244	<b>0.7623</b>
CNN-FCN	9.8688	0.3344	<b>11.324</b>	0.7566

TABLE VII. PERFORMANCE OF VARIOUS METHODS FOR FALL DATASET

Model	RMSE (\$/MWh)	NMSE	MAPE (%)	$R^2$
PERSISTENCE	17.314	0.5662	15.319	0.5907
ARIMA	15.851	0.5182	16.534	0.6108
RNN	14.947	0.4781	17.641	0.6290
LSTM	14.070	0.4525	15.980	0.6809
CNN-LSTM	<b>13.491</b>	<b>0.4305</b>	<b>14.771</b>	<b>0.7122</b>
CNN-FCN	15.215	0.5032	16.349	0.6526

As shown in Tables IV~VII, boldfaced digits of RMSE, NMSE, MAPE and  $R^2$  signify the best results. As can be seen in Table IV, CNN-FCN has the best values of RMSE, NMSE and  $R^2$  compared to other methods for the winter dataset. In Table V (spring result), LSTM attains the best results in terms of RMSE, NMSE and  $R^2$  while Table VI reveals that CNN-LSTM dominates the other methods in terms of NMSE and  $R^2$  for summer. As shown in Table VII, CNN-LSTM also obtains the best results by examining all performance metrics for the fall dataset.

Based on the above results, it shows that different characteristics of data in each season lead to various performances. In terms of  $R^2$ , CNN-LSTM is the best model having the greatest value in two seasons (summer and fall). When RMSE is used as a metric of performance, LSTM gains the lowest value in two (spring and summer) out of four seasons. All values of  $R^2$  in all four seasons were low because the data in four seasons have very high values of kurtosis and skewness indicating that data have a large number of outliers. In general, the deep learning-based LSTM, CNN-LSTM and CNN-FC outperform the others. In addition, performances of various methods are season-dependent.

## V. CONCLUSION

In this paper, comparative studies were conducted by statistic method, traditional supervised learning-based neural network and deep learning-based networks. Data preprocessing consisting of data grouping, analysis coefficients of variation, normality and stationarity were performed first. Then persistence, ARIMA, RNN, LSTM, CNN-LSTM and CNN-FCN were used to carry out locational marginal price forecasting. From the simulation results, it can be found the deep learning-based neural networks generally outperform the other methods. The persistence always gains the worst accuracy according to the performance metrics of RMSE, NMSE and  $R^2$ .

## ACKNOWLEDGEMENT

Authors thank the Ministry of Science and Technology in Taiwan for its support in MOST 109-3116-F-008-005 and 108-2221-E-033-023. No potential conflict of interest relevant to this article was reported.

## REFERENCES

- [1] K. Wang et al., "Wireless big data computing in smart grid", IEEE Wireless Communications, vol. 24, no. 2, pp. 58-64, 2017.
- [2] C. Xiang-ting, Z. Yu-hui, D. Wei, T. Jie-bin and G. Yu-xiao, "Design of intelligent demand side management system respond to varieties of factors", 2010, pp. 1-5.
- [3] L. Deng, Z. Li, H. Sun, Q. Guo, Y. Xu, R. Chen, J. Wang, "Generalized Locational Marginal Pricing in a Heat-and-Electricity-Integrated Market" IEEE Trans. on Smart Grid, vol. 10, no. 6, pp. 6414 - 6425, Nov. 2019.
- [4] L. Peterson, A. Nair and P. Ranganathan, "Short-term forecast for locational marginal pricing (LMP) data sets", 2018 North American Power Symposium, 2018. Available: 10.1109/naps.2018.8600581.
- [5] P. Kuo and C. Huang, "An electricity price forecasting model by hybrid structured deep neural networks", Sustainability, vol. 10, no. 4, p. 1280, 2018.
- [6] M. Cerjan, I. Krzelj, M. Vidak and M. Delimar, "A literature review with statistical analysis of electricity price forecasting methods," Eurocon 2013, Zagreb, Croatia, 1-4 July, 2013.
- [7] J. Nowotarski and R. Weron, "Recent advances in electricity price forecasting: A review of probabilistic forecasting," Renewable and

- Sustainable Energy Reviews, vol. 81, pp. 1548-1568, 2018. Available: 10.1016/j.rser.2017.05.234 [Accessed 15 March 2020].
- [8] J. Lago, F. De Ridder and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Applied Energy*, vol. 221, pp. 386-405, 2018.
- [9] S.S. Shapiro and M.B. Wilk, "An analysis of variance test for normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, p. 591-611, 1965.
- [10] D.A. Dickey and W.A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, p. 427-431, 1979.
- [11] K. Pearson, "IX. Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 216, no. 538-548, pp. 429-457, 1916.
- [12] G. Notton and C. Voyant, "Chapter 3: Forecasting of Intermittent Solar Energy Resource," *Advances in Renewable Energies and Power Technologies*, vol. 1, pp. 77-114, 2018.
- [13] Y. Hong and C. Hsiao, "Locational marginal price forecasting in deregulated electricity markets using artificial intelligence", *IEE Proceedings - Generation, Transmission and Distribution*, vol. 149, no. 5, pp. 621-626, 2002.
- [14] G. Box, G. Reinsel and G. Jenkins, *Time Series Analysis Forecasting and Control*, 3rd ed. Englewood Cliffs: Prentice Hall, 1994.
- [15] G.V. Houdt, C. Mosquera, G. Nápoles, "A review on the long short-term memory model. *Artif Intell Rev.*, 2020, <https://doi.org/10.1007/s10462-020-09838-1>.
- [16] P. Kuo and C. Huang, "An electricity price forecasting model by hybrid structured deep neural networks," *Sustainability*, vol. 10, no. 4, p. 1280, 2018.
- [17] Y. Bengio, Y. Lecun, "Convolutional networks for images, speech, and time-series," *The handbook of brain theory and neural networks*, pp. 255-258, Oct. 1998.
- [18] I. J. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.
- [19] P. Mandal, T. Senjyu, N. Urasaki, T. Funabashi, and A. K. Srivastava, "A novel approach to forecast electricity price for PJM using neural network and similar days method," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 2058-2065, 2007.
- [20] A. Motamedi, H. Zareipour, and W. D. Rosehart, "Electricity price and demand forecasting in smart grids," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 664-674, 2012.
- [21] A. Pourdayaei, H. Mokhlis, H. A. Ilias, S. H. A. Kaboli, S. Ahmad, and S. P. Ang, "Hybrid ANN and artificial cooperative search algorithm to forecast short-term electricity price in de-regulated electricity market," *IEEE Access*, vol. 7, pp. 125369-125386, 2019.

# Navigation Technology Using Inertial Elements to Compensate for GPS Signal-Shaded in Real Time

Guo-Shing Huang

*Institute of Electronic Engineering  
National Chin-Yi University of  
Technology  
Taichung City, Taiwan, R.O.C.  
hgs@ncut.edu.tw*

Po-Chun Hsu

*Institute of Electronic Engineering  
National Chin-Yi University of  
Technology  
Taichung City, Taiwan, R.O.C.  
conanfourom9@gmail.com*

Ming-Cheng Kao

*Department of Electronic Engineering  
Hsiuping University of Science and  
Technology  
Taichung City, Taiwan, R.O.C  
kmc@hust.edu.tw*

**Abstract**—Nowadays, the development of GPS (Global Positioning System) technology has become mature, and has begun to develop in the direction of high-precision positioning and navigation. Navigation data such as position, speed, and heading provided by GPS are very important in military or commercial fields. However, in addition to improving the accuracy of positioning, its reliability must also be considered. Therefore, RTK (Real Time Kinematic) differential positioning technology is used to perform real-time error correction and basic inertial components consist of gyroscope, accelerometer, and magnetic compass are used to compensate and improve reliability. In this study, a car is used as a vehicle to experiment the route near the campus. The collected XYZ position coordinate data of the non-difference, differential positioning, and original vehicle are calculated by MATLAB offline operation to obtain  $\lambda$ (longitude),  $\Phi$  (latitude),  $h$  (height), and the data sensed by integrated inertial elements in the GPS signal-shaded roads as auxiliary navigation, and then compare it with the existing and non-differential positioning methods to obtain a more optimized drawing path. The map achieves the compensation effect. In this study, the position of the shaded road is obtained by the second integration of the gyroscopic azimuth change and the heading angle of the magnetic compass through the accelerometer signal. The Extended Kalman Filter algorithm is used as its solution to estimate position, speed and heading of the vehicle in real time to achieve high accuracy and high reliability of real-time positioning and navigation.

**Keywords**—GPS, RTKDGPS, Extended Kalman Filter, Positioning, Navigation, Inertial Elements

## I. INTRODUCTION

GPS is now quite commonly used for navigation. There are many people who use positioning for delivery, logistics, and travel. It is an essential tool that is already closely related to life, in terms of logistics and travel. We often find ourselves in mountainous areas with poor reception or in sheltered areas such as tunnels, to compensate for this lost signal. F. Zhang et al. used Kalman filters to make predictions [1]. Therefore, this study uses inertial components to make compensation predictions.

The use of inertial components has become a fairly common sensor in life. It is often used to sense the attitude of the car body. The azimuth output of the gyroscope is used to sense the turning of the car body. The inertial system for this study was also used in a joint project with ADI OPTICS CO., LTD, to introduce a vehicle body attitude sensing system, which shows the commonality and practicality of this inertial system.

In the inertial compensation system, on the rugged road, especially on steep inclines, there is a sudden increase in cumulative error on steep inclines. Post-processing applications will use an extended Kalman filter algorithm versus a traditional compensation method. The difference between the two sides of the error in reducing inertial components can be seen. Gyroscopes, accelerometers, and magnetic compasses are an integral part of the inertial navigation system. In a sheltered environment, maybe it's not so precise. However, when corrected by the estimated measurements of the extended Kalman filter, this will reduce many errors, and the magnetic compass direction can be used to determine the section of the road to be compensated by measuring the acceleration with an accelerometer. The distance is calculated by integrating the points twice, and then use the GZ of the gyroscope, which is the angle of turn per second of heading, to calculate the forward angle. Lastly, with reference to the conversion of the latitude and longitude coordinates of Taiwan into meters, adding the error volume correction derived from the above gyroscopic GZ and the distance synthesis of the accelerometer's two integrals. The final compensation will be obtained in the road. This method of compensation is due to the fact that no GPS signal is use. K. M. Ng et al. studied that under different environmental influences, RTK positioning will change accordingly[2]. The roads may have little impact on large areas of the country, but there are buildings all around the road in Taiwan, and the roads are narrow, so this method can be used to compensate for the poor conditions inside the tunnel and outdoors. Consolidation of the above it can help most environments without GPS signal.

The extended Kalman filter in [3] is used to integrate INS and GPS signals into the solution, in this paper, the extended Kalman filter will be used to correct the gyroscope, acceleration gauge error, make the actual path more desirable for compensation.

## II. INERTIAL COMPENSATION SYSTEM AND ARCHITECTURE

The experimental data carrier itself averaged 8 km/h with an inertial element architecture such as Fig. 1. The sensor is obtained by the 6 axes gyroscope and the accelerometer of MPU6050. The magnetic compass is a QMC5883L chip. The gyroscope and accelerometer are essential sensing elements in this system. The compass is a biased auxiliary layer used to determine the orientation of the current load, combined with a triple sensor such as Fig. 1. The integrated system architecture of the hardware and GPS receiver is shown in Fig. 3.



Fig. 1. Hardware entity diagram of inertial components.



Fig. 2. Hardware architecture for integrating GPS and inertial component sensing systems on carriers.

The data obtained from the sensing is processed by MATLAB.  $A_x$  and  $A_y$  of the accelerometer are calculated from the vector trigonometric function to obtain the movement of the carrier and the declination of the gyroscope. The eq. (1) is used to derive the vector distance that  $A_x$  and  $A_y$ .

$$s = \frac{1}{2}at^2 \quad (1)$$

Fig. 3 Shows the declination angle  $G_z$  of the sensor gyroscope and the data of  $A_x$  and  $A_y$  of the accelerometer required to move the vehicle forward. The trigonometric function of a vector is used to derive the distance and direction of the forward motion.

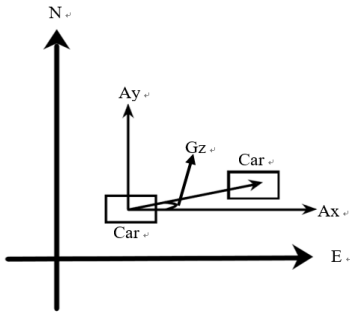


Fig. 3. Vehicle travel vector diagram.

The  $G_z$  is used to determines the angle of the carrier turns in Fig. 3. The trigonometric functions is used to calculate the synthetic vector of  $S_x$  and the synthetic vector of  $S_y$ . According to the above, we can see the actual distance forward.

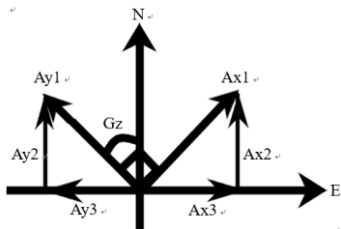


Fig. 4. Vector analysis.

In Fig. 4, we can see that the directions of the vector  $E$  axis forward. The equation (2) can be obtained from the subtraction difference of  $Ay_3$  and  $Ax_3$ . The equation (3) can be obtained from the summing of vector  $N$ -axis of  $Ay_2$ , which is in the same direction. Then the actually distance of the carrier moves can be obtained.

$$S_x = A_{x3} - A_{y3} \quad (2)$$

$$S_y = A_{x2} + A_{y2} \quad (3)$$

$A_{x3}$  is equal to the cos component of  $A_{x1}$ .  $A_{y3}$  is equal to the cos component of  $A_{y1}$  (4).  $A_{x2}$  is equal to the sin component of  $A_{x1}$ .  $A_{y2}$  is equal to the sin component of  $A_{y1}$  (5). The gyroscope  $G_z$  is used to sense the direction of turn of the vehicle. The direction of turn of the vehicle (2) and (3) are combined to get the forward position of (4) and (5).

$$S_x = A_{x1}\cos(G_z) - A_{y1}\sin(G_z) \quad (4)$$

$$S_y = A_{x1}\sin(G_z) + A_{y1}\cos(G_z) \quad (5)$$

### III. REAL-TIME DYNAMIC DIFFERENTIAL GPS SYSTEM

The common modern GPS [4] is a satellite navigation system that simply uses navigation signals for timing and ranging. The GPS system is divided into three sections: user, space and control. However, there are always many factors that can cause errors at the time of acceptance. As is clear from a comparison of Fig. 5 and Fig. 6. The general execution of GPS position uncorrected as shown in Fig. 5 with RTK-based dynamic real-time differential positioning correction techniques such as Fig. improves the accuracy of the position.

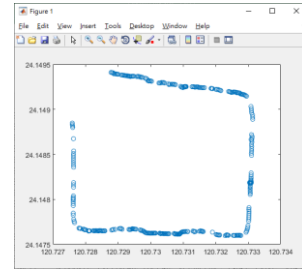


Fig. 5. Non-differential GPS obscured navigation path.

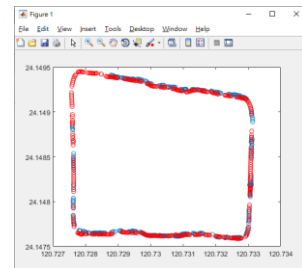


Fig. 6. Comparison of GPS navigation paths without differential and with differential.

In Fig. 7 you can see that there are a lot of floating positions in the height due to the lack of differential correction, with RTK's dynamic real-time corrected differential positioning technology. There is obviously a fairly smooth output pattern.



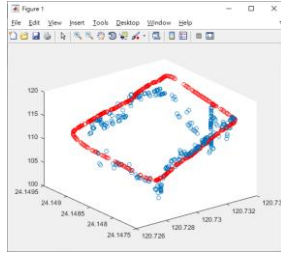


Fig. 7. Comparison of 3D navigation paths with or without differential GPS reception signals.

The so-called RTK uses a GNSS network composed of multiple base stations to evaluate the positioning error of the area covered by the base station [4]. The observation data from nearby physical base stations generate a Virtual Base Station (VBS) as an RTK. The base station is shown in Fig. 8.

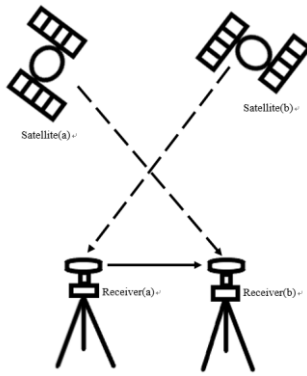


Fig. 8. RTK GPS positioning schematics.

#### IV. INERTIAL COMPONENT COMPENSATION SYSTEM PROPOSAL

First, integrate the development version with QMC5883L and MPU6050 as Fig. 9. The system block diagram of USB connection between development board and PC is shown in Fig. 10. The VCC input voltage of MPU6050 and QMC5883L development board is 3~5V, and PC connection baud rate is 9600Hz.

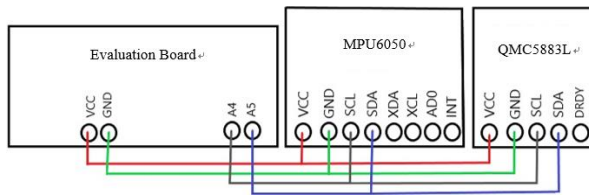


Fig. 9. Inertial component integration development version.

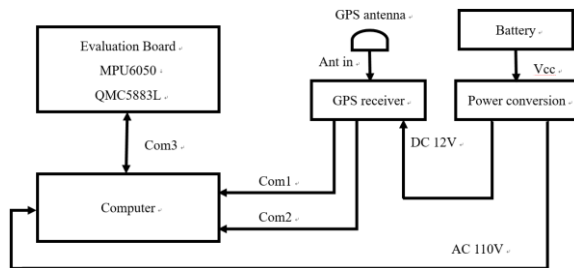


Fig. 10. Block diagram of integrated inertial components and GPS receiver system.

Parallel the x-axis of the carrier and the integrated development version to the longitudinal line, the y-axis is perpendicular to the vertical body line, inertial component sensing, i.e., the direction of travel and speed of the carrier itself, the combination of formulas (4) and (5) with (8) yields the position of the carrier itself.

#### V. EXTENDED KALMAN FILTER

To this day, the Kalman filter is still a popular algorithm. The most widely used of these is the extended Kalman filter. It is due to most systems are non-linear in nature. This algorithm can linearize a non-linear system, eliminate its noise to achieve the best possible prediction. In a joint project with ADI OPTICS CO., LTD, the extended Kalman filter is also suitable for use, with the vehicle attitude sensing system. The elimination of cumulative errors makes the vehicle attitude sensing system more accurate.

This paper examines when the GPS signal of a carrier is not received due to environmental factors, using inertial elements for compensation and extended Kalman filters commonly used in navigation, to estimate the best path for the sheltered section of the road.

##### A. Execution of Path Estimation Algorithm with Extended Kalman Filter

Extended Kalman filter combines gyroscope, accelerometer measurements, the acceleration sensing values of the accelerometer are used to estimate the position by quadratic integration. The actual movement of X and Y are synthesized by using the triangle function vector of Ax and Ay of the accelerometer. And then add up the movement position of the 2D floor plan of the carrier, but there would be a cumulative error, so the problem is solved by the extended Kalman filter.

The experimental steps are followed:

- The last point of latitude and longitude [E, N] before masking is established as the starting point, and then using the trigonometric function of the accelerometer and gyroscope to synthesize the components (4) and (5) to estimate the position of the next point. The vector of defined system state variables  $x = [E \ N \ Gz]^T$ , the observation vector,  $z = [\lambda \ \phi \ Gz]^T$ , the state equation that describes the system.

$$x(k+1) = f(x(k), k) \quad (6)$$

$$z(k+1) = h(x(k+1), k+1) \quad (7)$$

Each degree of latitude and longitude from the actual length of Taiwan is about 30.922 meters. The longitude is approximately 28.2084 meters, 1 degree equals 3600 seconds, so divide (4) by 30.922\*3600 and (5) by 28.2084\*3600 to get the movement per second. The change in latitude and longitude is based on the last point of the occlusion plus the change in latitude and longitude of movement per second. It is possible to predict the position of the next second.  $[E(k) \ N(k) \ Gz(k)]^T$  set to state variable matrix.  $[Sx/11319.2 \ Sy/101550.24 \ 1]^T$  set as transfer matrix.

$$\begin{bmatrix} E(k+1) \\ N(k+1) \\ Gz \end{bmatrix} = \begin{bmatrix} E(k) \\ N(k) \\ Gz \end{bmatrix} + \begin{bmatrix} Sx/11319.2 \\ Sy/101550.24 \\ 1 \end{bmatrix} \quad (8)$$

The equation (9) is the observation vector matrix, the definition of (7) can be set to the same output as (6).

- $[\lambda \ \phi \ Gz]^T$  set to measure the variable matrix.

$$\begin{bmatrix} \lambda \\ \phi \\ Gz \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} E(k+1) \\ N(k+1) \\ Gz \end{bmatrix} \quad (9)$$

### B. Procedure for Extended Kalman Filter Algorithm

Set Start Estimation State  $\hat{x}(0|0)$  and differential condition covariance matrix  $P_x(0|0)$  [5].

$$\hat{x}(0|0) = m_x(0) \text{ 及 } P_x(0|0) = 0 \quad (10)$$

- Prediction Status

$$\hat{x}(k+1|k) = f(\hat{x}(k|k), k) \quad (11)$$

- Status Linearization

$$f(x(k), k) = f(\hat{x}(k|k), k) + \phi(k+1, k)[x(k) - \hat{x}(k|k)] + \dots \quad (12)$$

- Differentiate  $f(x(k), k)$

$$\phi(k+1|k) = \left. \frac{\partial f(x(k), k)}{\partial x(k)} \right|_{\hat{x}(k|k)} \quad (13)$$

- Predicting the covariance matrix, this Q(k) is preset to 0 because it is not needed.

$$P_x(k+1|k) = \phi(k+1|k)P_x(k|k)\phi^T(k+1|k) + \Gamma(k+1, k)Q(k)\Gamma^T(k+1|k) \quad (14)$$

- $\hat{x}(k+1|k)$  linearization yields.

$$h(x(k+1), (k+1)) = h(\hat{x}(k+1|k), k+1) + H(k+1)[x(k+1) - \hat{x}(k+1|k)] + \dots \quad (15)$$

- Differentiation of  $x(k+1)$

$$H(k+1) = \left. \frac{\partial h(x(k+1), k+1)}{\partial x(k+1)} \right|_{\hat{x}(k+1|k)} \quad (16)$$

- Calculating k+1 for gain k, R(k+1) is set to 0 because it is not necessary here:

$$K(k+1) = P_x(k+1|k)H^T(k+1)[H(k+1)P_x(k+1|k)H^T(k+1) + R(k+1)]^{-1} \quad (17)$$

- Update (11), (14) using the gain K correction of (17)

$$\hat{x}(k+1|k+1) = \hat{x}(k+1|k) + K(k+1)[z(k+1) - \hat{z}(k+1|k)] \quad (18)$$

- Updating the covariance matrix of  $P_x(k+1|k)$ , I of (19) equals a matrix of size 1 of the same magnitude as the gain  $K(k+1)H(k+1)$ .

$$P_x(k+1|k+1) = [I - K(k+1)H(k+1)]P_x(k+1|k) \quad (19)$$

## VI. EXPERIMENTAL RESULTS AND DISCUSSION

Among the experimental results, it can be seen that the specified compensation for Fig. 1 was initially made in the conventional way using the inertial element gyroscope. Gz senses the turning direction of the carrier as Fig. 3 Gz is measured in degrees per second of angular change in the turn, will continue to grow in a cumulative manner. The unit of acceleration is 1 g/sec. The acceleration is given by the quadratic integral formula (1) to determine the distance traveled. The magnetic compass is used to determine the direction of the carrier's secondly travel in this experiment. In

order to judge the obscured section of road, the fusion of the gyroscope The changepe and accelerometer components are achieved by calculating the path and direction of travel of the carrier in the equations in (4) and (5).

Experiments are carried out using traditional methods of compensation, a compensation was granted in the portion of the original lower right corner of Fig. 5. However, there are still many errors in the traditional compensation sections in Fig. 11. A comparison of the overlap in Fig. 13 can be seen. In this one of the more obvious comparisons, the conventional calculation of compensation can only marginally converge with the original road.

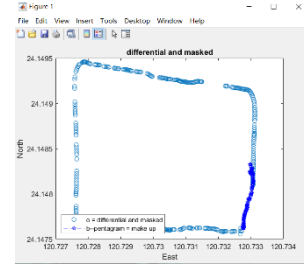


Fig. 11. Conventional compensation with differential GPS signal masking inertia element.

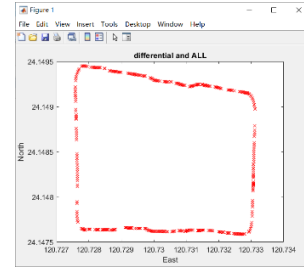


Fig. 12. Whole-turn path with differential GPS signal navigation.

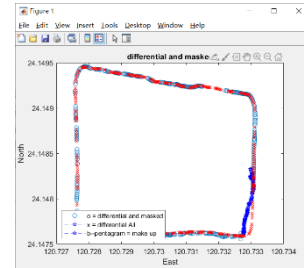


Fig. 13. Comparison of conventional compensation for inertial elements with and without GPS signal masking.

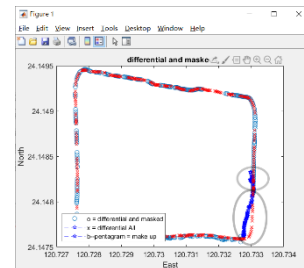


Fig. 14. Conventional compensation error for GPS signal blocking inertia element.

From the above-mentioned experiments, it is known that traditional path compensation has the disadvantage of cumulative error as Fig. 11. Therefore another way of

reducing the cumulative error in the experiment was to use an extended Kalman filter. The cumulative error can be decreased to complete the best path compensation estimation by the linearization of the filter reduces. The difference in the experimental effect of masking compensation can be seen by comparing Fig. 11 with Fig. 15. Fig.16 shows the local magnification comparison of Fig.15 using the Extension Kalman Filter Algorithm.

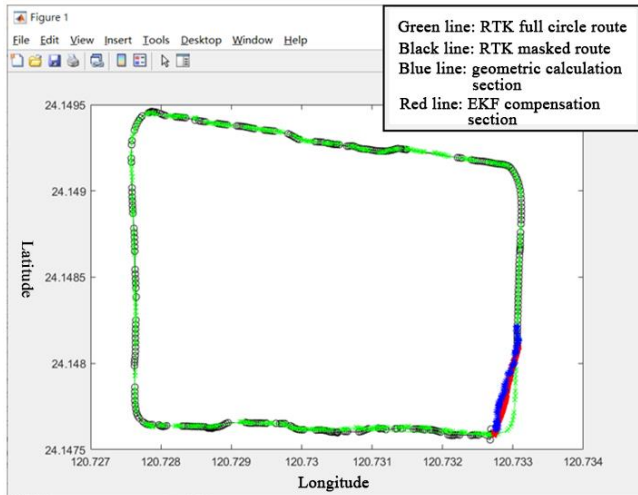


Fig. 15. GPS signal blocking inertia element compensation prediction using the EKF method.

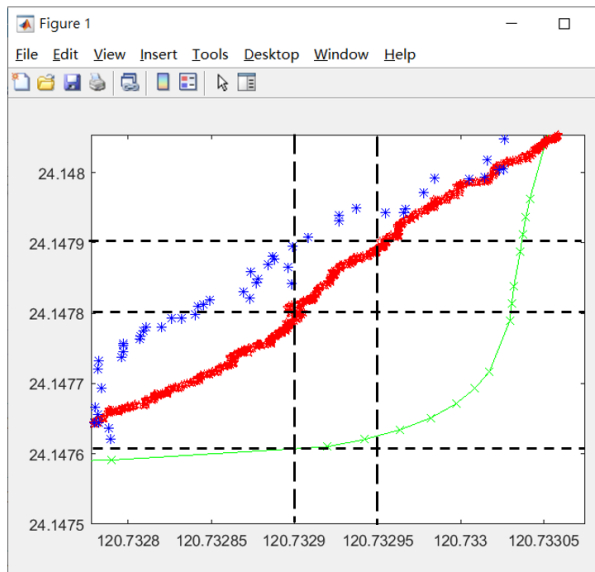


Fig. 16. Local magnification comparison shows the red line with the EKF method.

In experimental results, the real-time dynamic differential GPS system shows the differential and non-differential heights in Fig. 13. There are many drifts without differential, that simply using GPS without the assistance of other systems. There will be many errors by differential correction, and with compensation correction by the extended Kalman filter method in Fig. 15. The resulting compensation path is much smoother and more stable.

## VII. CONCLUSION

In this result, we can continuously display the navigation coordinates on the navigation maps when the GPS signal is temporarily lost due to masking or geo-environmental factors. It exhibited near-precise coordinates effectively by basic inertia components. The inertial components of gyroscope, accelerometer, and magnetic compass are used to determine the angle of turn per second, position per second, and direction of travel of the carrier. The coordinates of the 2D map can be calculated with conversion to compensate for the loss of GPS signal on the path of the original vehicle. This study is a real-time dynamic differential positioning technique using RTK (Real Time Kinematic) to use the inertial components for real-time compensation for auxiliary navigation. In order to achieve the position of the vehicle on the map, the GPS signal will not be obscured by the coordinates, and the navigation display will be continuous.

## ACKNOWLEDGMENT

Authors are very grateful to Ministry of Science and Technology, Taiwan for the financial support. (MOST 108-2221-E-167-027-MY2)

## REFERENCES

- [1] F. Zhang, B. Shan, Y. Wang, Y. A. Hu, Z. Guo, and H. Teng, "MIMU/GPS integrated navigation filtering algorithm under the condition of satellite missing," 2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC), 02 March 2020.
- [2] K. M. Ng, J. Johari, S. A. C. Abdullah, A. Ahmad, and B. N. Laja, "Performance evaluation of the RTK-GNSS navigating under different landscape," 2018 18th International Conference on Control, Automation and Systems (ICCAS), pp. 1424-1428, 13 December 2018.
- [3] W. Zhou, X. W. Qiao, F. Meng, and H. Zhang, "Study on SINS/GPS tightly integrated navigation based on adaptive extended Kalman filter," The 2010 IEEE International Conference on Information and Automation, pp. 2344-2347, 19 July 2010.
- [4] J. -C. Juang, "Satellite navigation," December 2012, Chuan Hwa Publishing Ltd., Taiwan.
- [5] G.-S. Huang, C. -K. Tung, P. -S. Hung, S. -R. Huang, and S. -T. Yan, "Using DR algorithm to integrate GPS and inertial elements," Proceedings of the 2002 China Geographic Information Society Annual Meeting and Academic Symposium, pp. B-42, Fengjia University Geographic Information System Research Center, Taichung City, 3-4 October 1991.

# An Automatic Approach for Estimation of CPR Signal using Thoracic Impedance

Van-Truong Pham

School of Electrical Engineering  
Hanoi University of Science and Technology  
Hanoi, Vietnam  
truong.phamvan@hust.edu.vn

Thi-Thao Tran ✉

School of Electrical Engineering  
Hanoi University of Science and Technology  
Hanoi, Vietnam  
thao.tranthi@hust.edu.vn

**Abstract**— Quality of chest compressions is an important measure with regards to cardiopulmonary resuscitation (CPR). For chest compression quality assessment, it is important to detect or estimate the chest compressions in heart rhythms. In this study, we proposed a new approach for estimation of the CPR during out-of-hospital cardiac arrest, and assessment of chest compression quality via CPR parameters. The CPR signal is estimated by combining candidate modes derived from an ensemble empirical mode decomposition (EEMD) based on frequencies of modes and thoracic impedance signal. The CPR parameters including compression numbers, compression rates, flow-time, and no-flow-time, from the estimated CPR are also evaluated with parameters derived from the reference CPR. Experiments show good agreements with high correlation between parameters by estimated and reference CRP signals, that demonstrate the performances of the proposed approach.

**Keywords**—Cardiopulmonary resuscitation, Cardiac arrest, Thoracic impedance, Adaptive Filter, Empirical mode decomposition.)

## I. INTRODUCTION

The importance of chest compressions during cardiopulmonary resuscitation (CPR) has been emphasized in resuscitation guidelines. It has been proved that the chest compressions play an important role in the treatment of cardiac arrest. The most recent guidelines recommend that chest compressions should be provided with a depth of at least 5 cm at a rate between 100 to 120 compressions per minute (cpm), to allow full chest recoil with minimum interruptions in compressions [1]. To estimate the CPR artifact and assess the quality of chest compression, a variety number of approaches have been introduced in the literature, such as the method of using accelerometers in [2, 3], automated external defibrillators (AEDs) in [4], frequency of compressions in [5], and thoracic impedance (TI) signal in [4-6]. Among the above approaches, the method of using TI signal as the reference of an adaptive filter is of high interest. However, the adaptive filter has shortcomings associated with filter parameter setting. In addition, the performance of the adaptive filter is sensitive to the reference signal, so that if the reference like TI signal is of high fluctuation, the estimated CPR is not reliable. In another approach, Lo et al. in [7] combined dominant modes from the empirical mode decomposition (EMD) then reconstruct a reference signal for a least mean square (LSM) adaptive filter. The output of the adaptive filter is finally used to estimate the CPR artifact.

This method possesses advantages of EMD in handling nonstationary signals with different time scales. Nevertheless, selecting the dominant modes is nontrivial since the order of the mode might vary with input signals. Besides, the number of modes changes accordance with signal waveform, so choosing a certain order of mode might not be robust. In addition, the approach of direct estimation of CPR from corrupted ECG has the drawback that the CPR might be estimated (non-zero) even when the chest compression (true CPR) is not existed in the duration of considered signal.

In this study, inspired from the EMD and the idea of spectrum subtraction, we proposed a new approach to estimate the CPR fluctuations using corrupted ECG and thoracic impedance signals, without using adaptive filter. In the mode decomposition, we use the Ensemble Empirical Mode Decomposition (EEMD) to address the mode-mixing problem in EMD. We then present an automatic algorithm to estimate the chest compressions using the ECG acquired from defibrillators and the thoracic impedance (TI) signals. In more detail, we applied the EEMD to decompose the ECG signal into modes, and computed the frequency of each mode, as well as the frequency of TI signal. Then we compared the mode frequency with the TI frequency. The modes whose frequencies close to TI frequency are then added. The added signal is assigned as the estimated CPR signal. Based on the presence or absence of CPR in the duration of the considered signal, the flow-time or no-flow-time can be identified, and compression parameters are calculated.

## II. BACKGROUND

### A. Cardiac arrest and chest compression

Cardiac arrest refers to the abrupt loss of heart function and can result in death if the heart suddenly stops working properly. Cardiac arrest is caused by irregular rhythms of heart which are called arrhythmias. Arrhythmias are typically analyzed by automated external defibrillators (AED) to shockable or un-shockable rhythm in which the un-shockable rhythm can be treated by CPR.

According to resuscitation guidelines [8], cardiac arrhythmias can be classified in to following types: Asystole, ventricular fibrillation, ventricular fibrillation and Pulseless electrical activity [9]. Asystole (AS) is the state rhythm in which there is no cardiac electrical activity. Ventricular

fibrillation (VF) is disorderly and irregular electrical activity in the heart's ventricles that causes the heart to beat quickly and out of rhythm. Ventricular tachycardia (VT) is a type of regular, fast heart rate that arises from improper electrical activity in the ventricles. Pulseless electrical activity (PEA) refers to cardiac arrest in which there is an organized electrical activity but there is no pulse.

Rhythm classification is normally based on the analysis of biomedical signals. Along with the ECG signal, other signals recorded by external defibrillators such as chest compression depth related acceleration or pressure signal, and thoracic impedance signals are also widely used. Pressure signal is normally collected from the sensor system that is fitted on an extra pad of the defibrillator. The information is delivered based on the sensitivity of sensor to the chest movement [10]. The chest compressions are acquired when pressing the chest from 4 to 5 cm that is necessary for the vital organ with blood circulation. It has been reported that the typical compression rate should be in the range of 100 -120 compressions per minute (cpm). Correlated with the pressure signal, the thoracic impedance signal is associated with the electrical impedance of biological tissues. Following the Ohm's law, a voltage drop can be measured when passing a current through the tissues [10].

### B. Fourier transform

Fourier transform is most amenable to analytical insights and manipulations of analog systems. The two types of Fourier transform, continuous-time Fourier transform, and the discrete Fourier transform (DFT) are considered as respectively transforms for continuous-time and discrete-time duration signals. Among the two types of Fourier transform, the DFT is widely used in digital computations with the use of a fast algorithm called the fast Fourier transform (FFT) [11] for computing the DFT. The equations respectively expressed in the forward and inverse transforms of the Fourier transform is given as:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \quad (1)$$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega \quad (2)$$

The DFT is used in discrete time signal,  $x[0], x[1], \dots, x[N-1]$ . The DFT  $X[0], X[1], \dots, X[N-1]$  is discrete frequency and also finite duration. The DFT is expressed in the forward and inverse forms as:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-jk\omega_0 n} \quad k = 0, 1, \dots, N-1 \quad (3)$$

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{jk\omega_0 n} \quad k = 0, 1, \dots, N-1 \quad (4)$$

$$\text{with } \omega_0 = \frac{2\pi}{N}$$

### C. Ensemble empirical mode decomposition

Empirical Mode Decomposition (EMD) [12] is an adaptive time-space analysis method that is suitable for processing of non-linear and nonstationary signals and has been widely applied in many fields such as signal processing [7], medical imaging [13], seismic wave [12]. The EMD can be compared to other signal analysis methods like Fourier

Transforms and wavelet decomposition. The EMD decomposes any given data into a set of finite number of narrow band modes, also known as intrinsic mode functions, which are derived directly from the data. Meanwhile Fourier and wavelet transforms incorporate predefined fixed basis for signal modelling and analysis.

The EMD decompose a signal  $x(t)$  into a set of sub-signals, namely modes or intrinsic mode functions (IMFs). To be considered as a mode, a signal must satisfy two conditions: (1) the number of extrema and the number of zero-crossings must be equal or differ at most by one, and (2) at any point, the mean values of the upper and lower envelope are zeros. The EMD extracts the input signal into  $N$  modes expressed as the following equation

$$x(t) = \sum_{m=1}^N c_m(t) + r_N(t) \quad (5)$$

where  $c_m(t)$  is the  $m^{\text{th}}$  mode ( $m=1, \dots, N$ ), and  $r_N(t)$  is the residue of the decomposition process.

Though EMD has achieved performances in signal analysis, it often suffers from mode mixing problem. To overcome the mode mixing issue, the ensemble empirical mode decomposition has been proposed by Wu and Huang [14]. EEMD is a noise assisted data analysis method that can separate scales naturally without any a priori subjective criterion selection as in the original EMD algorithm. EEMD consists of sifting an ensemble of white noise-added signal, and can be briefly described as following steps:

- Denote  $w^i[n]$  ( $i=1, \dots, M$ ) as different realizations of white Gaussian noise, and then generate:  $x^i[n] = x[n] + w^i[n]$
- Decompose each  $x^i[n]$  by EMD to get their modes  $IMF_k^i[n]$ , with  $k=1, \dots, K$  indicate the modes
- Take the average of the corresponding  $IMF_k^i$  as:

$$IMF_k^i = \frac{1}{M} \sum IMF_k^i[n]$$

## III. METHODOLOGY

### A. Signal preprocessing

The input signals for analysis of cardiac arrhythmias include ECG, thoracic impedance (TI), and pressure signals. The ECG signals contaminated with CPR artifact were recorded by defibrillators with a sampling rate of 250 Hz. The TI signals are used for estimating the frequency of the EEMD modes to reconstruct the CPR signal, and also eliminating the artifacts that are contaminated in the ECG signals. The pressure signals, or compression depths, are used as the reference to assess the performance of the proposed CPR assessment method. For efficient computation, the TI and reference signals are resampled to the frequency of 250Hz. All signals are divided into five-second epochs, and then preprocessed by a filtering step using a 4<sup>th</sup> order bandpass Butterworth filter. The filtering frequency range is set to 0.5-30Hz for ECG signals, while the range for the TI and reference signals is 0.7-5Hz. For CPR quality assessment, each one-minute segment was analyzed.

### B. Performing EEMD for ECG signal

The ECG signals filtered by the bandpass filter are decomposed into a set of finite number of narrow band modes, also known as IMFs, by the EEMD method [14]. The number of IMFs is based on certain epoch of the signal. After decomposing the ECG signal, we computed the frequency of



each mode using fast Fourier transform. The spectrograms of the modes are also obtained. For demonstration, we showed an example of EEMD results for a ten-second epoch of ECG signal in Fig.1. In Fig.1 a, the plots of the input ECG signal epoch and its decomposed modes in time-series are shown. The frequencies obtained from the frequency distributions by performing FFT on the input signal and modes are also given in the titles of the corresponding subplots. Besides, the Gabor spectrograms of input ECG signal together with its IMFs are also provided in Fig. 1b. As can be seen in Fig. 1b, by performing EEMD, we obtained the narrow band IMFs with different frequencies that can be separated from the frequency range of the input signal.

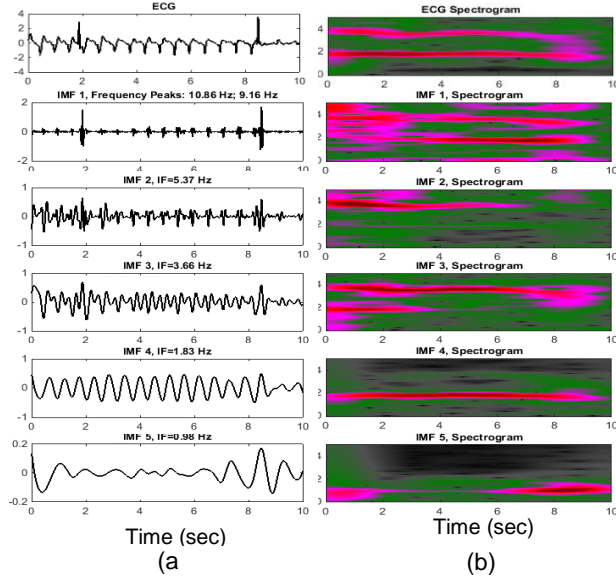


Fig.1. Illustration of EEMD decomposing input ECG signal epoch into modes (IMFs). (a) ECG signal and its modes decomposed by EEMD, and (b) corresponding Gabor spectrograms of ECG signal and the modes.

### C. Relations between modes and acquired signals

It has been validated by previous works [5, 6], that the frequency of TI signal is related to the chest compression depth. Therefore, to estimate the frequency of CPR signal, we can use the frequency of the TI signal. The relation between the frequency of CPR and TI signals can be demonstrated by the experiment in Fig.2. For this experiment, we first applied the FFT on the input ECG signals as well as the TI and reference CPR signals to find their frequency distributions. From the frequency distributions, we computed the frequencies of those signals. As can be observed in Fig. 2, the frequencies of TI and the reference CPR signals are quite close, with the values of 1.76 Hz and 1.83 Hz, respectively. It is noted that the value of 1.83 Hz in CPR signal is equivalent to 110 compressions per minute (cpm) for this epoch. We also can see that the value of 1.83 Hz is also one of two peaks in the frequency distribution of ECG signal that was originally contaminated by CPR artifact. It is worth mentioning that the input ECG signal epoch used in the experiment for Fig. 2 is also the ECG shown in the top of Fig. 1. a. When relating the frequencies of decomposed modes in Fig.1 and the signal frequencies in Fig. 2, we can see that the frequency of IMF 4 is close to the frequency of TI signal, and also coincided with the frequency of reference CPR.

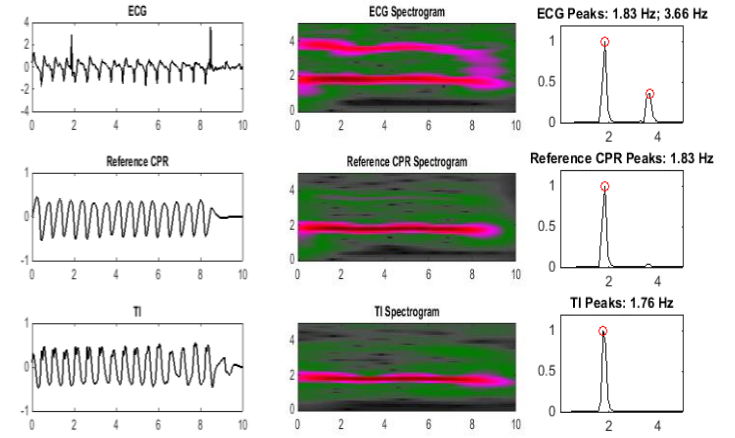


Fig. 2. Plots in time series (left), spectrograms (middle), and frequency distributions with peaks (right) of the input ECG (first row), reference CPR (second row), and TI signals (third row).

### D. The proposed approach for CPR estimation

Inspired by the relations between the ECG signal as well as its decomposed IMFs with reference CPR and TI signals, we propose a new approach for CPR estimation. The main idea in this study is to estimate the CPR signal from the decomposed modes from ECG signals via performing EEMD. The proposed approach is stemmed from the fact that the ECG signal is generally contaminated with CPR artifact that presents almost periodic waveforms. On the other hand, the decomposed modes are normally also in periodic waveforms, so it is reasonable to estimate the CPR signal from the modes. In fact, the compression depths, used for reference CPR signals are rare and hard to acquire, so it is valuable to build a good algorithm for estimating the CPR. In this study, we use the TI signal to estimate the CPR signal since it is normally available while the compression signals are usually not available in acquisition systems.

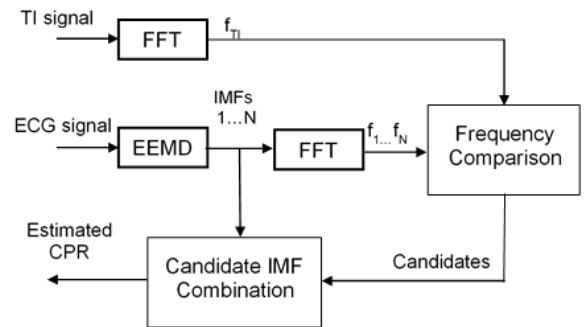


Fig. 3 The proposed algorithm for CPR signal estimation

The main steps of the proposed algorithm for estimating CPR signal is shown in Fig. 3. First, we decompose the ECG signal epoch into IMFs then compute their frequencies,  $f_1, f_2, \dots, f_N$ . The frequency of TI signal,  $f_{TI}$ , of that epoch is also calculated. The frequencies are computed by using FFT. Next, we compare the frequency set of modes with TI frequency, to check whether the frequency of the mode coincides with TI frequency or not. In other words, we select the modes whose frequency close to  $f_{TI}$ , with a tolerance chosen as 0.2 Hz in this study. The selected modes are considered as candidate IMFs, and then combined for CPR signal estimation. It is noted that, if no candidate IMF is

found in an epoch, the estimated CPR signal for the epoch will be zero.

#### E. The proposed method for CPR quality assessment

In this study, to assess the CPR quality, we compute parameters associated with the quality of chest compression that were defined in [1, 15]. The parameter factors include the no-flow-time (NFT), flow-time (FT), compression number (CN), and chest compression rate (CCR). No-flow-time refers to a pause of more than 1.5 seconds in chest compressions. Flow-time is defined as the total CPR duration minus no-flow-time. Compression number is calculated as the number of compressions in the considered segment (i.e., one-minute segment). The chest compression rate is computed as chest compression number divided by flow-time.

To evaluate the performance of the proposed approach for CPR estimation, we compare the parameters computed from the estimated CPR, called automatic, and those by reference CPR, called reference hereafter. For each segment, the flow-time/no-flow-time in the reference CPR is manually determined by viewing the presence/absence in the reference CPR signal. To estimate the CPR parameters, the flow-time/no-flow-time is automatically determined based on the numbers of epochs without CPR identified by the proposed candidate mode combination step. The number of compressions, also called compression numbers, are computed based on the number of frequency peaks above a threshold (i.e., 50% of the average amplitude) on the CPR signal in time series. The compression rates are then defined based on the corresponding compression number and flow-time parameters.

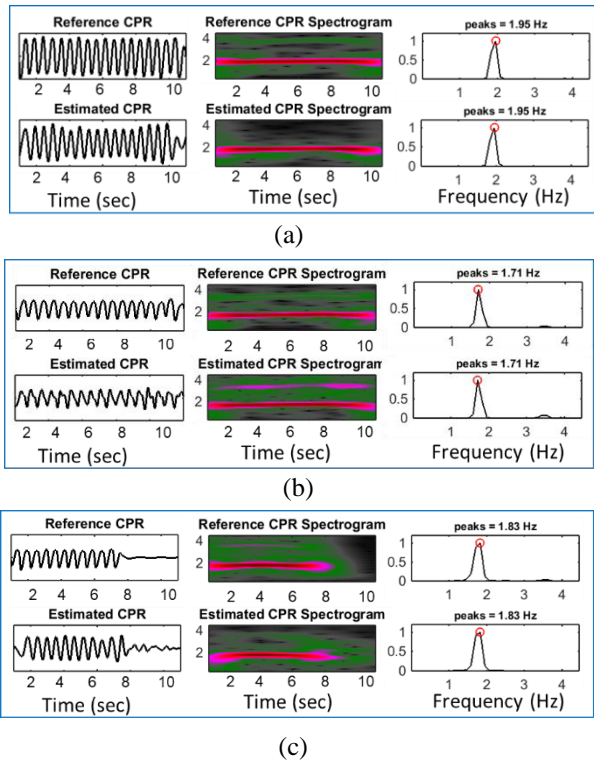


Fig. 4. Representative example of the estimated CPR in comparison with the reference CPR signals in time series (left), Gabor spectrograms (middle), and Frequency distributions with peaks (right) in patients groups of (a) AS, (b) VFVT, and (c) PEA.

## IV. RESULTS

For CPR signal estimation and CPR quality assessment, we applied the proposed algorithms to the database with patient groups including AS (686 segments), VFVT (1076 segments), and PEA (778 segments). The signal processing steps including bandpass filtering, EEMD, FFT, and CPR analysis are implemented using Matlab.

#### A. Evaluation of estimated CPR signals

This experiment shows the agreement between the estimated CPR signal with the true CPR signal in terms of waveform and frequency. A representative case is shown in Fig. 4 for all patient groups. As can be seen from the first two columns of this figure, the estimated CPR signal are in good agreement with the reference CPR signal for both time series and spectrogram plots. Moreover, as can be seen in Fig. 4c, the frequency distributions of the two signals in the corresponding group are similar, with the same frequency.

#### B. CPR quality assessment by the proposed approach

The CPR signals estimated by the proposed approach are then used for calculating parameters including no-flow-time, flow-time, compression number, and compression rate. The parameters of the estimated CPR signals, called automatic CPR, are also compared with those derived from the reference CPR signals.

TABLE I. CPR PARAMETERS DERIVED FROM AUTOMATIC (ESTIMATED CPR) AND THE REFERENCE CPR. ABBREVIATION: FT :FLOW-TIME ; NFT: NO-FLOW-TIME; CN: COMPRESSION NUMBER; CR: COMPRESSION RATE; AUTO :AUTOMATIC; REF: REFERENCE. VALUES ARE IN MEAN (STD)

		FT (s)	NFT (s)	CN (cpm)	CR (1/min)
AS	Auto	47.0 (15.3)	11.9 (15.2)	85.6 (29.1)	101.9 (26.2)
	Ref	47.4 (16.1)	12.1 (16.1)	85.6 (29.9)	101.7 (27.6)
VFVT	Auto	42.8 (17.6)	16.7 (17.6)	78.3 (36.5)	100.0 (37.5)
	Ref	41.1 (19.9)	18.4 (19.9)	77.3 (38.7)	102.0 (37.1)
PEA	Auto	42.5 (18.4)	17.0 (18.3)	76.7 (34.7)	98.9 (33.8)
	Ref	41.7 (19.8)	17.8 (19.8)	76.4 (36.6)	97.4 (36.7)

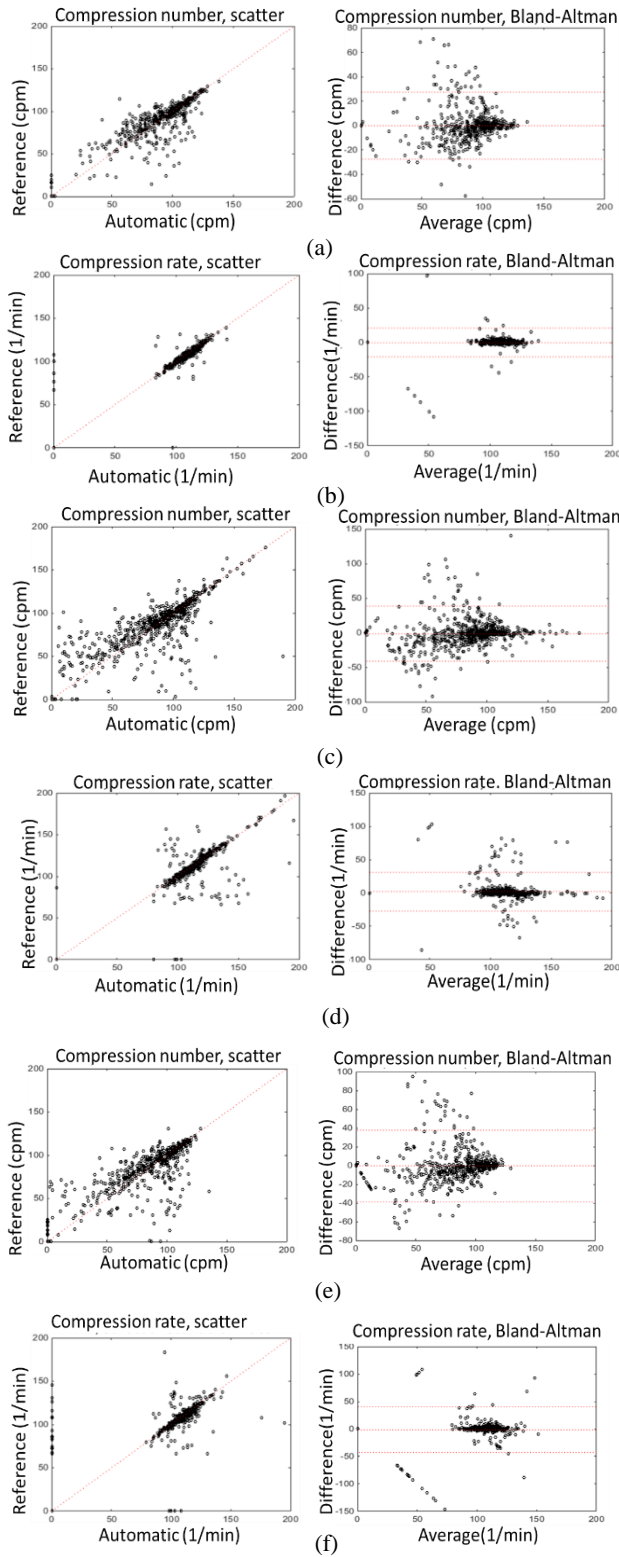


Fig. 5. The scatter plots and the Bland-Altman mean-difference plots of CPR parameters including compression numbers and compression rates. The plots show the correlations between the automatic and the reference CPR signals for patient groups. (a) Compression number of AS, (b) Compression rate of AS ;(c) Compression number of VFVT, (d) Compression rate of VFVT; (e) Compression number of PEA, (f) Compression rate of PEA.

For quantitative evaluation, the mean and standard deviation (STD) values of CPR parameters for both automatic and reference signals are provided in Table 1. The parameters are calculated for every one-minute segment of data from patient groups with AS, VFVT, and PEA. The table showed that all parameters derived by the automatic are close with those by reference CPR signals.

For a more evaluation, we also provided the correlations of compression number and compression rate parameters derived from the automatic and reference CPR signals. The evaluation is depicted by scatter as well as the Bland-Altman mean-difference plots in Fig. 5 for patient groups including AS, VFVT, and PEA. Besides, as shown in the plots, parameters computed from estimated CPR are highly correlated with those calculated from reference CPR signals.

### C. Method comparison

To validate the performance of the proposed approach, we reimplemented the approach by Lo et al. [7] and applied for the database. The correlation coefficients by the method in comparison with the proposed algorithm for compression number and compression rate are provided in Table 2. As presented in the table, the proposed method obtained better correlation coefficient values for both factor parameters, compression number and compression rate. The obtained parameters are with high correlation coefficients, above 86% for compression number for all groups and 90% for compression rates in AS and VFVT groups.

TABLE II. COMPARISON OF CPR QUALITY ASSESSMENT BY CORRELATION COEFFICIENTS FOR COMPRESSION NUMBER (CN), AND COMPRESSION RATE (CR) PARAMETERS

Methods		Lo et al.	Proposed
AS	CN	0.71	0.89
	CR	0.88	0.93
VFVT	CN	0.70	0.86
	CR	0.91	0.92
PEA	CN	0.73	0.86
	CR	0.82	0.83

### V. CONCLUSIONS

This study has presented an approach for assessment of the CPR quality for analyzing ECG signals. Using the ECG signals acquired from defibrillators and thoracic impedance signals, we propose to estimate the CPR signal. We first perform the EEMD on the input ECG signal to separate the ECG signal into modes, then utilize the information from frequency of thoracic impedance to select candidate modes. The selected modes are then combined to be the estimated CPR signal. The estimated CPR signals are then applied on

the proposed CPR quality assessment algorithm to compute CPR parameters. The proposed approach has been applied for the database including patient groups of Asystole, ventricular fibrillation/ventricular tachycardia, and pulseless electrical activity. Experiments show good agreements and high correlation between parameters computed from estimated and reference CPR signals that show the performances of the proposed approach.

#### ACKNOWLEDGMENT

This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2020-PC-017.

#### References

- [1] J. Kramer-Johansen, D. Edelson, H. Losert, K. Köhler, and B. Abella, "Uniform reporting of measured quality of cardiopulmonary resuscitation (CPR)," *Resuscitation*, vol. 74, pp. 406-417, 2007.
- [2] B. Abella, J. Alvarado, H. Myklebust, D. Edelson, A. Barry, N. O'Hearn, *et al.*, "Quality of cardiopulmonary resuscitation during in-hospital cardiac arrest," *JAMA*, vol. 293, pp. 305-310, 2005.
- [3] B. Abella, N. Sandbo, P. Vassilatos, J. Alvarado, N. O'Hearn, H. Wigder, *et al.*, "Chest compression rates during cardiopulmonary resuscitation are suboptimal: a prospective study during in-hospital cardiac arrest," *Circulation*, vol. 111, pp. 428-434, 2005.
- [4] T. Valenzuela, K. Kern, L. Clark, R. Berg, M. Berg, D. Berg, *et al.*, "Interruptions of chest compressions during emergency medical systems resuscitation," *Circulation*, vol. 112, pp. 1259-1265, 2005.
- [5] U. Irusta, Ruiz J, de Gauna. SR., T. Eftestøl, and J. Kramer-Johansen, "A least mean-square filter for the estimation of the cardiopulmonary resuscitation artifact based on the frequency of the compressions," *IEEE Trans Biomed Eng.*, vol. 56, pp. 1052-1062, 2009.
- [6] U. Ayala, T. Eftestøl, E. Alonso, U. Irusta, E. Aramendi, S. Wali, *et al.*, "Automatic detection of chest compressions for the assessment of CPR-quality parameters," *Resuscitation*, vol. 85, pp. 957-963, 2014.
- [7] Lo MT., Lin LY., Hsieh WH., Ko PC., Liu YB., Lin C., *et al.*, "A new method to estimate the amplitude spectrum analysis of ventricular fibrillation during cardiopulmonary resuscitation," *Resuscitation*, vol. 84, pp. 1505-1511, 2013.
- [8] J. Soar, *et al.*, "European resuscitation council guidelines for resuscitation 2015, ECG-Based Classification of Resuscitation Cardiac Rhythms for Retrospective Data Analysis," *Resuscitation*, vol. 95, pp. 100-147, 2015.
- [9] W. Skjeflo, T. Nordseth, J. Loennechen, D. Bergum, and E. Skogvoll, "ECG changes during resuscitation of patients with initial pulseless electrical activity are associated with return of spontaneous circulation," *Resuscitation*, vol. 127, pp. 31-36, 2018.
- [10] Q. Tan, G. Freeman, F. Geheb, and J. Bisera, "Electrocardiographic analysis during uninterrupted cardiopulmonary resuscitation," *Crit Care Med.*, vol. 36, pp. S409-S412, 2008.
- [11] G. D. Bergland, "A guided tour of the fast Fourier transform," *IEEE Spectrum*, vol. 6, pp. 41-52, 1969.
- [12] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, E. H. Shih, Q. Zheng, *et al.*, "The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis," *Proc. R. Soc. Lond.*, vol. 454A, pp. 903-995, 1998.
- [13] TT Tran., VT Pham., C Lin., HW Yang., YH Wang., KK Shyu., *et al.*, "Empirical Mode Decomposition and Monogenic Signal based Approach for Quantification of Myocardial Infarction from MR Images," *IEEE J. Biomed. Health Informat.*, p. DOI 10.1109/JBHI.2018.2821675, 2018.
- [14] Z. Wu, N. E. Huang, and X. Chen, "The multi-dimensional ensemble empirical mode decomposition method," *Adv. Adapt. Data Anal.*, vol. 1, pp. 339-372., 2009.
- [15] L. Lin, M. Lo, W. Chiang, C. Lin, P. Ko, K. Hsiung, *et al.*, "A new way to analyze resuscitation quality by reviewing automatic external defibrillator data," *Resuscitation*, vol. 83, pp. 171-176, 2012.



# A Study on Patterns of Neural Activity Generation from A Bio-realistic Cerebellum Neural Network

Vo Nhu Thanh\*, Pham Anh Duc, Le Hoai Nam, Dang Phuoc Vinh, Tran Ngoc Hai

Faculty of Mechanical Engineering

The University of Da Nang-University of Science and Technology

54 Nguyen Luong Bang, Lien Chieu, Da Nang, Viet Nam

\*Corresponding author: vnthanh@dut.udn.vn

**Abstract**— The biological neural network signal flow is a complex process and difficult to interpret. The purpose of this study is to build a model of a simplified cerebellum neural network and investigate its working behavior in generating neuronal patterns. The model is constructed by referring to the neural anatomic structure of the mammal cerebellum and then simulated using Matlab Simulink software. The artificial neural network dimension is much smaller than the size of the real cerebellar neural network in mammal due to hardware simulation resource limitation. Thus, the utilization of extensive-scale spiking neural networks is limited due to the grown of computational costs associated with propagating bio-realistic neural models. The parameters correlated with the neuron function also affect the biological plausibility of the spiking neural network. At the conclusion, we present the pattern generation of a cerebellum-like neural network and briefly explained why such patterns are generated.

**Keywords**— Biological Neural Network, Neural Pattern, Matlab Simulink, Neural Anatomic Structure

## I. INTRODUCTION

The animals collect data of the surrounding environment through their receptive organs. Then, they process this data and select proper actions. All of the schemes transpire in their neurotic system. Billions of neurons are correlated and interacted with each other to process information from receptive input and then to carry the action controls to the motor system [1]. A single neuron averagely forwards its signals out to over 10.000 others [2]. Thus, the biological neural network signal flow is a complex process and difficult to interpret.

Thus, getting the knowledge and then assembling a biological neural network have been a very attractive topic for scientists since the discovery of a neuron behavior in 1932 [1]. Through many investigations and studies, the neural network has evolved into the 3rd generation also known as spiking neural network.

The first generation of artificial neural networks [3] was a very simple paradigm. A neuron produced a digital ‘high’ signal if the total amount of input signals was larger than a threshold value. Notwithstanding their simplistic formation and its digital production, the first generation neural networks have been strongly implemented in many complex systems. For example, in digital computations, any Boolean function could be concocted by a multilayer perceptron with a single hidden layer. The 50 years old first generation of artificial neural networks has become pretty old technology and rarely applied today.

Neurons of the next generation of the artificial neural network used a consecutive activation function to perform

their signals. Sigmoid and hyperbolic tangent was the two ordinarily used activation functions in the second generation. Feed-forward and recurrent neural networks were conventional models of the second-generation neural networks. These models were more dominant than their predecessors because they could operate properly among any analog system, making them suitable for analog calculations. In the case these models were implemented with a threshold function, they would also work very well for digital computations and yet with fewer neurons within the network as the first generation [4].

Biological neurons convey data messages by utilizing tiny and prompt raises in charge. The natural neural signals are commonly recognized as action potentials or spikes. The neurons encode data messages in both their firing frequency and the timing of individual spikes. Spiking neural networks recognized as the third generation of the neural network are more persuasive and biologically plausible than the non-spiking forerunners for the transient data are encoded within the neuron signals. Nevertheless, it is more complex to model and analyze a spiking neural network.

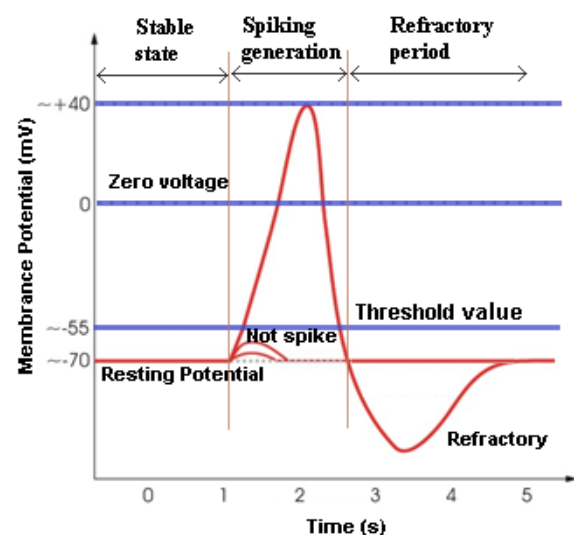


Fig. 1. A Biological neural signal.

Each membrane has two significant levels of potential which are the resting potential which is around  $-70$  millivolts (mV) and the threshold potential which is around  $-55$  mV. Resting potential is the level of the membrane potential at its stable state while threshold potential is the level at which a neuron will excite a spike. Signal inputs or synaptic inputs to a neuron make the membrane potential to rise or fall, which



are respectively called depolarize or hyperpolarize. Action potentials are produced if the total amount of depolarization increases the membrane potential beyond the threshold value. When an action potential transpires, the voltage on the membrane rapidly rises to a particular value approximately 40mv and quickly drops down. The membrane's potential decreases beneath the resting level, and slowly recovers to the resting potential lately. The refractory period is the duration when the voltage on the membrane is less than the resting potential. Hence, a standard action potential contains three phases which are depolarization, hyperpolarization, and refraction. The entire process takes roughly 4-5 milliseconds for most neurons. The whole process is presented in Figure 1.

The applications of the spiking neural network are similar to regular artificial neural networks. Nevertheless higher bio-realistic properties neural networks are usually employed as the investigation tool for analyzing the mechanism of the nervous scheme. For instance, based on a proposal on the anatomical arrangement of a biological neuronal circuitry and its function, a model of a spiking neural network is built. This model is then computer-simulated to get the output signals. The output signals of this model are then correlated with the electrophysiological recordings of the actual biological neural network circuit for validation and determining the plausibility of the prediction model as the study of the short timing function of the cerebellar neural network by Thanh et.al. [5]. Nevertheless, these models gradually rely on the evolution of computational capability. The utilization of extensive-scale spiking neural networks is limited due to the grown of computational costs associated with propagating bio-realistic neural models. The parameter correlated with the neuron function also affects the biological plausibility of the spiking neural network. In this study, we present the patterns of neural activity generated by an artificial cerebellum-like neural network and briefly explained why such patterns are generated.

## II. SPIKING NEURON MODEL

Louis Fapicque proposed the earliest model of spiking neural network which was known as the integrated and fire model as described in equation 1. The membrane voltage increased with time if an input current was applied until it reached a constant threshold  $V_{th}$ . Then, a refractory period was followed (equation 2) to restore the membrane potential to its steady-state [1], [6].

$$I_c = C_m \frac{dV_m(t)}{dt} \quad (1)$$

$$f(I) = \frac{I}{C_m V_{th} + I t_{ref}} \quad (2)$$

Hodgkin and Huxley introduced the first systematic model of a spiking neuron that illustrated how biological neural signals or spikes were launched and generated based on the investigation about giant squid neural responses in 1952[7]. They modeled a nerve cell as an electrical component. Their model is explained in equations (3) to (5).

The membrane voltage is " $V_m$ ". The lipid bilayer of the membrane is defined as a capacitance " $C$ ." A voltage-gated ion channel is recognized as electrical conductance " $g_i$ " and its value varies relying on electric charge and timing. The electrical conductance of the leakage ion channel is labeled as " $g_L$ ". Voltage sources " $E_n$ " are the electrochemical gradients,

and its value is restricted by the concentrations of the ionic varieties. " $I_p$ " is the current of the ion. The equilibrium potential (voltage at steady state) of the  $i^{th}$  ion channel is " $V_i$ ".

The current flowing through a cell's membrane is:

$$f(I) = \frac{I}{C_m V_{th} + I t_{ref}} \quad (3)$$

Current through a given ion channel is:

$$I_i = g(V_m - V_i) \quad (4)$$

The total current through the membrane is given by:

$$I = C_m \frac{dV_m}{dt} + g_k(V_m - V_k) + g_{Na}(V_m - V_{Na}) + g_L(V_m - V_L) \quad (5)$$

The most major feature of a neural network is its learning mechanism. Similar to the earlier generations, the learning of spiking neural networks also based on synaptic weights adaptation over time. By changing the synaptic weights, the neural signals progress within a neural network is modified; hence, the signals conveyed to a neuron are adjusted. As a consequence, the output signals of the neural network are also modified. This is a fundamental mechanism of learning, which is known as synaptic plasticity in neuroscience [7].

Two main types of plasticity are depression or potentiation which either weaken or strengthen the input signal. Regarding duration, short-term synaptic plasticity happens in about tens of milliseconds to a few minutes, while long-term plasticity continues from several minutes to a few hours. The NMDA and AMPA glutamate receptors are essentially involved in molecular behaviors of synaptic plasticity. The opening of NMDA channels appears a rise in postsynaptic transmitting capability, which is linked to long-term potentiation (LTP); while NMDA ion channels lower the transmitting capability of post-synaptic, which linked to long-term depression (LTD). These synaptic plasticities are critical in learning and retention of the nervous system [8].

## III. CEREBELLUM-LIKE NEURAL NETWORK

The main purpose of this section is to construct a bio realistic neural network model by referring to the structure of the cerebellum. Lewis and Miall have published a study on the contribution of the cerebellum role in precision and accurate timing control of muscle actuator in 2003[9]. Thus, we construct a model that could adaptively update its timing and gain prediction based on the anatomic structure cerebellum in this study. Figure 2 is a model structure of the cerebellar interconnecting circuit of neurons.

The main types of neurons in the cerebellum include Purkinje cell (PKJ), Stellate cell (STL), Basket cell (BSK), Granule cell (GR), Golgi cell (GO), and Deep cerebellar nuclei (DCN). The signals from receptors to neurons and neurons to neurons are conducted by Mossy fibers (MFs), Climbing fibers (CFs), and Parallel fibers (PFs). The positive excitations are conducted from the MFs and CFs which then be synapsed directly onto the PJK. MFs conduct positive excitatory signals to GRs. The GRs give excitatory signals to PKJs through PFs. PKJs give negative inhibitory signals to DCN which delivery the output neural signal of the cerebellar cortex. STL, BSK, GO also give negative inhibitory signals to

other neurons include PKJ and GR. The random projection model proposed by Yamazaki and Tanaka is referred to as the standard structure to build the cerebellum model in this study[10]. Due to the small influence of inhibitory signals from BSK and STL to PKJ, we ignore these connections to save hardware resources. The timing control reception of the random projection model is primarily obtained from the granular layer. Thus, the timing encoded within the cerebellum output signal is miscarried if this layer has defects. The spiking neural network is built and modeled using Matlab Simulink software to obtain the neurons' activities pattern.

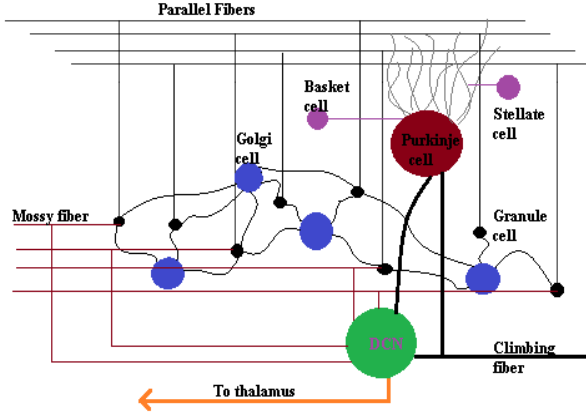


Fig. 2. Overview of the cerebellar circuit

#### IV. NEURON MODEL IMPLEMENTATION

The mathematical neuron model in this study uses the conductance-based leaky integrate and fire neuron model as shown in equations (6) to (8).

$$C \frac{dV}{dt} = -g_{leak}(V(t) - E_{leak}) - g_{ex:short}(t)(V(t) - E_{ex}) - g_{ex:long}(t)(V(t) - E_{ex}) - g_{inh}(t)(V(t) - E_{inh}) - g_{ahp}(t - \hat{t})(V(t) - E_{ahp}) + I_{External} \quad (6)$$

$$g_i(t) = \bar{g}_i \sum_j w_j \int_{-\infty}^t \alpha(t-s) \delta_j(s) ds \quad (7)$$

$$g_{ahp}(t - \hat{t}) = \exp\left(-\frac{(t-\hat{t})}{\tau_{ahp}}\right) \quad (8)$$

- $V(t)$  is the membrane potential of a neuron at time  $t$ .
- $C$  is the capacitance.
- $I_{External}$  is the external input current (for example from Mossy fiber).
- $E$  is the reversal potential.
- $g_{leak}(t)$  is the conductance function of leakage current.
- $g_{ex:short}(t)$  is the short-time excitatory conductance function.
- $g_{ex:long}(t)$  is the long-time excitatory conductance function.
- $g_{inh}(t)$  is the inhibitory input conductance function.
- $g_{ahp}(t)$  is an after-hyperpolarization input conductance function.
- $\bar{g}_i$  is the maximum conductance of  $i$  ( $i = [ex\_short, ex\_long, leak, ahp, inh]$ ).
- $w_j$  is the synaptic weight.

- $\delta_j(t)$  is the presynaptic neuron  $j$  action at time  $t$  (1 or 0).
- $\tau_{ahp}$  is the value of the after-hyperpolarization time constant.
- $\alpha(t)$  is the alpha function of a neuron, different type of alpha function is shown in Table 1.

The parameters of short-time and long-time excitatory signals are varied and affect the learning mechanism of the simplified cerebellum neural network. ( $a1$  varies from 5~10,  $a2$  varies from 1~2,  $a3$  varies from 20~80,  $a4$  varies from 5~15,  $a5$  varies from 10~50). For this study, we only showed the simulation result of fixed parameters  $a1=5$ ;  $a2=1$ ;  $a3=50$ ;  $a4=10$ ;  $a5=20$ .

TABLE 1. ALPHA FUNCTION

Neuron	Alpha function
PKJ	$\alpha_{exsht} = e^{-\frac{t}{a1}}$
GR	$\alpha_{exsht} = e^{-\frac{t}{a2}}$ $\alpha_{exlg} = e^{-\frac{t}{a3}}$ $\alpha_{inh} = \frac{7}{16}e^{-\frac{t}{7}} + \frac{9}{16}e^{-\frac{t}{59}}$
GO	$\alpha_{exsht} = e^{-\frac{t}{1.5}}$ $\alpha_{exlg} = \frac{3}{8}e^{-\frac{t}{31}} + \frac{5}{8}e^{-\frac{t}{170}}$
DCN	$\alpha_{exsht} = e^{-\frac{t}{a4}}$ $\alpha_{exlg} = e^{-\frac{t}{a5}}$ $\alpha_{inh} = e^{-\frac{t}{43}}$
IO	$\alpha_{exsht} = e^{-\frac{t}{10}}$ $\alpha_{inh} = e^{-\frac{t}{10}}$

The cerebellar cortex is a very big neural network but moderately simple and repeatedly arrangement with clearly defined input/output terminals [10]. The cerebellar cortex has three main layers which are an outer molecular layer, a Purkinje cell layer, and a granular layer. The outer molecular consists of cerebellar neurons; the Purkinje cell layer mainly consists of PKJ, and the granular layer is constructed from GRs and GOs [11]. The most synaptic inputs in the cerebellum are projected to the Purkinje cell layer to generate the output signals of the cerebellar cortex.

We feed three inputs consist of 2 30 Hz Poisson signals to GR clusters, and 1 pre-processed signal varies from 5 to 200 Hz to initialize the neural network through the Inferior olivary nucleus (IO) to simulate the working behavior of this model. The output from DCN specifies the neurons' activity pattern encoded the timing information of the neural network. This output is stored at the Matlab workspace for further analysis.

The alpha functions are presented as the exponential function; thus, depending on the time constant coefficients, the process can be fast or slow. In fact, the coefficients can be adaptively changed in different situations for different types of learnings and neural processing. The neurons fixed parameters are obtained from the other reliable literal researches [5,6] and presented in table 2.

TABLE 2. NEURON PARAMETERS

Parameter	Unit	Neuron type				
		PKJ	GR	GO	DCN	IO
$C$	pF	107	3	28	122	10
$g_{leak}$	nS	2.5	0.45	2.5	1.75	0.625
$E_{leak}$	mV	-68	-58	-55	-56	-60
$g_{exshort}$	nS	0.75	0.175	45	50	1
$E_{exshort}$	mV	0	0	0	0	0
$E_{exlong}$	nS	-	0.025	30	26	-
$E_{exlong}$	mV	-	0	0	0	-
$g_{inh}$	nS	-	0.025	-	30	0.175
$g_{ahp}$	nS	0.125	1	20	50	1
$E_{inh}$	mV	-	-82	-	-88	-75
$E_{ahp}$	mV	-70	-82	-73	-70	-75
$T_{ahp}$	ms	5	5	5	2.5	10
Threshold	mV	-55	-35	-52	-39	-50

By referring to the anatomic structure of the cerebellum as indicated in Figure 2, our cerebellum model is constructed for simulation using Matlab Simulink is shown in Figure 3. The model consists of 80000 GRs and 800 GOs divided into 800 GR clusters with 100 GR cells per cluster and 80 GO clusters with 10 GO cells per cluster. For ensuring the randomness of neuron projections, we divided GO clusters and GR cluster into 2 different pathways and recurrent one another. Between the pathways are the Signal splitter block and the Distribution block which are used to distributed signals from other neurons to the same neuron similarly to the biological neural network. The Initialize block is used to feed the transient signal to make sure all neurons are fired at the very beginning of the process. GRs receive positive input from MFs and negative inhibitory inputs from several nearby GO cells. The synaptic weight from MFs to GRs is 0.18 and from GOs to GRs is 2.0 recurrently with a probability of 2.5%. Each GO cell only gets the random positive input signals from nearly 100 surrounded GRs with a synaptic weight of 0.02 and a possibility of 5%.

The PKJs receive the positive excitation from many GRs through PFs with a small synaptic weight of 0.0058 and a probability of 7.5%. These projections from PFs to PKJs will ensure a firing rate of 90 to 100 signals/s at the beginning phase for PKJs. To prevent DCN from continuously firing and let its signal rapidly decreasing, the inhibitory synaptic weight of PKJs projecting to DCN is 0.5. Inhibitory synaptic from DCN to IO must prevent it from over-training; thus, this weight is selected to be 45. We only pay attention to the pattern generation mechanism of the cerebellum neural network, which we hypothesize to be the internal function of the GR layer. To save hardware resources and with the assumption of the GR layer effect on timing learning, we ignore other types of neuron such as BSK, STL, and unipolar brush cells in the calculation process.

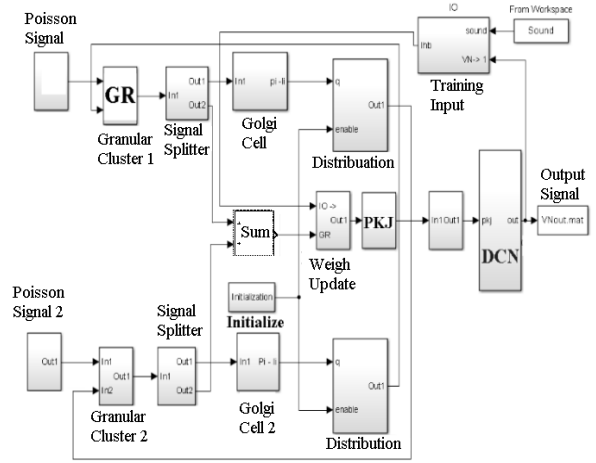


Fig. 3. Neuron model implementation in Matlab Simulink Software

## V. NEURAL SPIKING PATTERN ANALYSIS AND CONCLUSION

We run the simulations on Matlab Simulink to analyze and verify the behavior of the spiking neural network. We input the external signal with a transient portion of 200Hz for 1 millisecond as the initialization stage. Then, this signal frequency is switched to 5Hz and combined with a 30Hz Poisson signal lasting for the remaining period of 999 milliseconds simulation. The total time for simulating this neural network using a core i7-2.8 GHz, 16 Gb ram PC is about 3 hours.

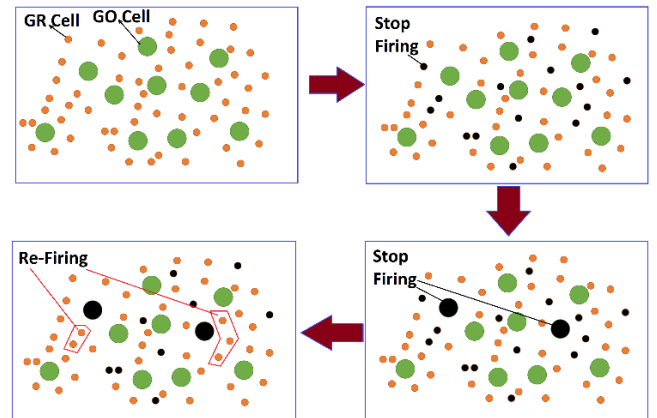


Fig. 4. Neuron activities of 100 random Granular Cell s

The mechanism to generate such a unique pattern is illustrated subsequently. At first, all GRs rapidly firing due to a high-frequency input of 200Hz transient signal. Next, all GRs excitatory signals are forwarded to GOs through a distribution block. The randomness from the distribution block ensures that GOs have various levels of firing rate. As mentioned above, GRs receive negative inhibitory inputs from several nearby GOs. Therefore; GRs that received strong inhibitory input from GOs likely stop firing. As a result of the non-spiking action of the GRs, some GOs that connect to inactivated GRs would have a lower firing rate. Obviously, the negative inputs from these GOs to its nearby GRs are also weakened. This is the condition for some of the GRs which were not active before can now produce an action potential. This looping sequences continuously occur at the GR layer, and distinctive patterns of GRs is produced. The process is clearly illustrated in Figure 4.

The results in Figures 5 and 6 indicate that a specific GR cell produces a unique neuron activity pattern. Some neurons firing at the start of the period and then stop; some neurons firing at the end of the simulation; and some neurons firing at the middle of the simulation period, while some other neurons firing at the beginning and the end of the simulation.

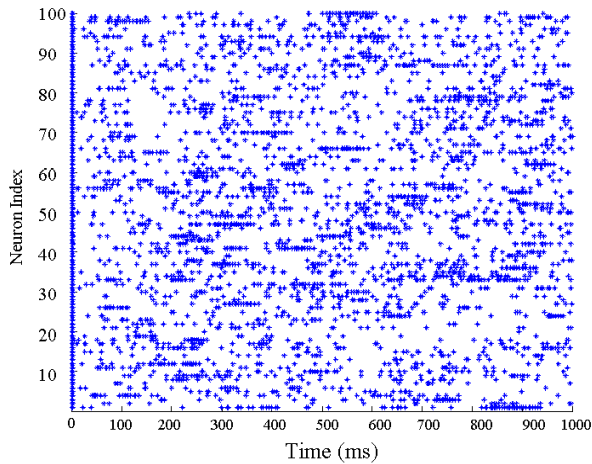


Fig. 5. Neuron activities of 100 random Granular Cells

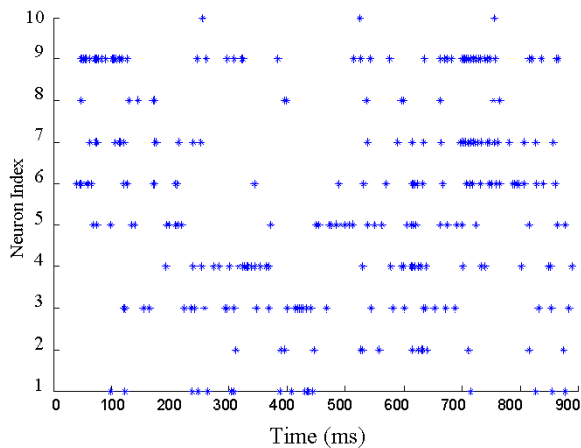


Fig. 6. Neuron activities of 10 random Granular Cells

The authors successfully constructed a bio-realistic neural network modeling the cerebellum to analyze and investigate its neural pattern generation mechanism. We are able to

explain why such patterns are generated as indicated in section V. The results of the spiking pattern shown in Figures 5 and 6 indicate the suitability of the study and this is fundamental for further explanation of the cerebellar function to mammalian behavior. This is an important contribution to intelligent mechanical systems for its usability. Since we only performed the simulation in Matlab software, the simulation time is very long, roughly around 3 hours. We would like to implement this cerebellum neural network to a high and parallel processing system such as GPUs or FPGA to increase the speed of calculation for real-time application in the next study.

#### ACKNOWLEDGMENT

This work was supported by The University of Danang-University of Science and Technology, code number of Project: T2020-02-02

#### REFERENCES

- [1] J. Vreeken, Spiking neural networks, an introduction, Technical Report UU-CS-2003-008 Institute for Information and Computing Sciences, Utrecht University, 1-5 (2002).
- [2] D. Bullock, J.C. Fiala, and S. Grossberg, A neural model of timed response learning in the cerebellum, *Neural Network* (7), 1101–1114 (1994).
- [3] R. FitzHugh, Impulses and physiological states in theoretical models of nerve membrane, *Biophysical Journal*, Vol. 1, 445–466 (1961).
- [4] D. Sterratt, B. Graham and A. Gillies, *Principles of Computational Modelling in Neuroscience*, Cambridge University Press (2011).
- [5] V.N. Thanh and H. Sawada, A Spiking Neural Network for Short-range Timing Function of a Robotic Speaking System, *The 3rd International Conference Proceedings on Control, Automation and Robotics (ICCAR 2017)*, pp. 184-187, Nagoya (2017).
- [6] V.N. Thanh and H. Sawada, Simplified Cerebellum-like Spiking Neural Network as Short-Range Timing Function for the Talking Robot, *Connection Science* (30,4), pp. 388-408 (2018).
- [7] W. Gerstner and W.M. Kistler, *Spiking Neuron Models. Single Neurons, Populations, Plasticity*, Cambridge University Press (2002).
- [8] R. Apps, M. Garwicz, "Anatomical and physiological foundations of cerebellar information processing," *Nature Reviews Neuroscience*, Vol. 6, pp. 297-311 (2005).
- [9] P.A., Lewis and R.C. Miall, Distinct systems for automatic and cognitively controlled time measurement: evidence from neuro imaging, *Current Opinion in Neurobiology* (13), pp. 250-255 (2003).
- [10] T. Yamazaki and S. Tanaka, A spiking network model for passage-of-time representation in the cerebellum, *The European Journal of Neuroscience* (26, 8), pp. 2279–2292 (2007).
- [11] Sander M. Bohte, Joost N. Kok, Applications of Spiking Neural Networks, *Journal of Information Processing Letters* (95, 6), pp. 519-520 (2005).

# Integration of MicroSCADA SYS600 9.4 into Distribution Automation System

The Khanh Truong

*The University of Danang-University of  
Science and Technology  
Danang, Vietnam  
khanhtruongnk93@gmail.com,*

Kim Hung Le

*The University of Danang-University of  
Science and Technology  
Danang, Vietnam  
lekimhung@dut.udn.vn,*

Minh Quan Duong\*

*The University of Danang-University of  
Science and Technology  
Danang, Vietnam  
dmquan@dut.udn.vn*

Tue Truong-Bach

*Department of Science Technology and Environment  
The University of Danang  
Danang, Vietnam  
tbtue@ac.udn.vn*

Van Phuong Vo

*Load Dispatch Division  
Danang Power Company  
Danang, Vietnam  
phuongvv@cpc.vn*

**Abstract—** In this paper, the Distribution Automatic System (DAS) is introduced and analyzed to improve the operation of power systems in Quang Nam province, Vietnam. The application of DAS contributes to quickly detect and isolate incidents, immediately restore the normal operation of the system and improve the power system reliability. The DAS is an effective support tool for the dispatchers in the remote operation, control and management of power systems. Previous publishes have done the research and application of DAS technology to the distribution grid but did not analyze of busbar segment problem and the programming for the DAS. This paper investigates the application of DAS using MicroSCADA SYS600 9.4 Pro software in the construction of smart grids in Quang Nam province and examines specific cases of faults on two Quang Nam feeders. Also, a program to implement the simulation of DAS is processed.

**Keywords—** *Distribution Automatic System, distribution grids, MicroSCADA SYS600 9.4 Pro, SCADA systems, Smart grid.*

## I. INTRODUCTION

The economy of Quang Nam province, Vietnam has been making particularly important changes, especially the industry, tourisms and services are developing rapidly. Therefore, the demand to meet power capacity, uninterruptible power supply and quality of electricity for the production processes and daily-life activities are increasing day by day. In addition, the development of renewable energy sources in Quang Nam, especially photovoltaic rooftop systems, poses challenges affecting the utility grid when these sources are grid-connected. The connection of these systems to the distribution grid in random order (depending on the location of the installed household) will cause phase imbalance and large difference in transmitted power if rooftop systems only focus on connecting to a single phase. Thus, the integration of automatic monitoring and operation technologies is one of the optimal solutions to meet the strict requirements in the energy development context of Quang Nam province.

The emergence of Distribution Automatic System (DAS) technology to the distribution grid based on the monitoring, management, control and data acquisition system of Supervisory Control and Data Acquisition (SCADA) will contribute greatly to quickly detecting reliability of power supply to customers and strengthen power quality.

The Distribution Automatic System – DAS, is a system that automatically controls the operation mode of the distribution grid to detect faulty elements and detach them from the system. The corrupted segment will be quickly isolated fragmented, and the power supply of the other elements will be restored in order to avoid power outage over a wide area.

The combination of DAS technology and SCADA system platform will bring positive effects to the Vietnamese power system and it will support the dispatchers in the process of system operation and management.

Despite of the previous studies about the application of the DAS, the relevant issues of the experimental DAS still have not been explored, such as the analysis of incidents on the grid or processes to build the DAS. Therefore, this paper will present the technology of the DAS on SCADA monitoring, control and data collection system applied to the distribution grid of Quang Nam province, Vietnam. Incident cases tested on the DAS, particularly two feeders of the 110kV Hoi An substation in Quang Nam province, will be highlighted on the analysis and research to provide appropriate solutions to detect and eliminate the incidents. In addition, the research will build database and control program for the system to execute system operation using MicroSCADA SYS600 9.4 Pro software and make recommendations to complete an applicable DAS for the Vietnamese power system.

## II. SCADA SYSTEMS

### A. Introduction to SCADA systems

SCADA (Supervisory Control and Data Acquisition) is an industrial automation management system with the function of controlling, monitoring and collecting data of the system. Thanks to the SCADA system, operators can identify and control the operation of electrical equipment through computers and communication networks. In the management and operation of the power system, the SCADA system plays a very important role in assisting operators in accurately tracking, monitoring and processing data in the power system [1].

The advantage of SCADA systems is to provide accurate and timely data that allows to optimize the operation of the system and process. Moreover, the power system using SCADA is always more efficient, reliable and safer.



The main components of a basic SCADA system are supervisory computers (also called supervisory center), remote terminal units (RTUs) and communication infrastructure. The RTUs are located at the substation to collect data as well as to monitor and send control signals to the central SCADA system via communication infrastructure (fiber-optic cable or 3G). The switchgears such as reclosers, load break switches (LBS) or ring main unit (RMU) have control panels that are connected to the SCADA system using 3G technology or fiber-optic cable channels, usually via 3G transmission lines using 3G modems [1-2].

### B. Principle and function of SCADA systems

- Principle of SCADA systems:

According to the predetermined period (about a few seconds), the host computer of the SCADA system at the supervisory center will perform the signal transmission for sequential scanning of substations, reclosers and LBS. These elements are equipped with RTUs or Gateway devices that allow the supervisory center to control devices through them [3].

- Function of SCADA systems:

The SCADA system has three main functions: monitoring, control and data collecting. In addition, the SCADA system also has the function of analyzing, processing and storing data; power flow calculation; short-circuit and reliability calculation; load demand management and providing database for other purposes [4-5].

## III. DISTRIBUTION AUTOMATIC SYSTEM

### A. Implementation of Distribution Automatic System

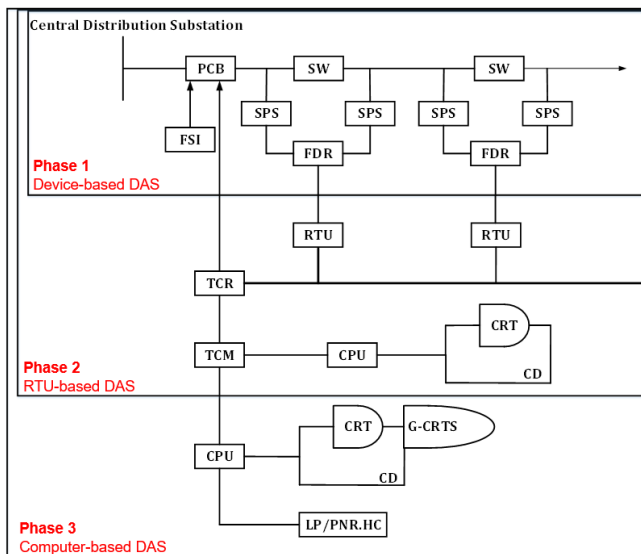


Fig. 1. Implementation phases of the DAS

According to Figure 1 the implementation of the DAS usually goes through 3 phases as follows:

- Phase 1: Installation of automatic circuit breakers (CB) and fault detecting relays (FDR) on medium voltage power lines. In phase 1, the incident area is automatically isolated from the system by devices on the medium voltage power line without activating devices at the supervisory center.

- Phase 2: Additional installation of RTUs and communication lines to receive information at the position of automatic CB in medium voltage power lines. At the supervisory center, remote control units and computer systems are installed to display the medium voltage grid in a simple form. Based on the information remotely obtained from the grid, the operators at the supervisory center will control automatic CB to isolate the faulty part by using the computers.
- Phase 3: Upgrading the functions of phase 2. At the supervisory center, supercomputers are used to manage the operation of the medium voltage distribution grid displayed according to the geographical map and adjust the process calculation automatically. After the completion of the above 3 phases, the distribution grid is completely remote-monitored [6-7].

### B. Communication protocols

Currently, most SCADA systems of the Vietnamese power system use IEC 60870-5-101 protocol for communication from control points to the SCADA system. Basically, IEC 60870-5-101 protocol meets the requirements of real-time measurement and control signals for control objects. However, with the characteristics of connection by serial communication interface, IEC 60870-5-101 protocol has many limitations in establishing physical communication channels, and it is difficult to expand the connection points on the system. With the development of communication protocols based on TCP/IP communication protocol, IEC 60870-5-104 protocol was appeared to become a communication solution for SCADA systems, creating many advantages in implementation as well as high stability in communication [8-9].

The IEC 60870-5-104 protocol was released by the International Electrotechnical Commission in 2000. The IEC 60870-5-104 protocol creates physical connections based on TCP/IP protocol so the implementation of communication on physical layers becomes simpler and is easily compatible with Gateway devices and RTUs of different companies.

Communication signal of IEC 60870-5-104 connecting from RTUs to SCADA system is carried out on Fast Ethernet (FE) physical layers of communication devices, or through E1/FE converters (main line). Backup line is also proposed to be made via low-cost Internet connections (3G/GPRS or ADSL) [10-11].

Some basic advantages of IEC 60870-5-104 communication protocol evaluated through the testing process are listed as follows:

- The IEC 60870-5-104 protocol is obviously compatible with the IEC 60870-5-101 protocol for the link layer and the application layer. Therefore, the database construction for control objects on the MicroSCADA system does not change.
- The IEC 60870-5-104 supports an interface connection using Ethernet (FE ports), so the expense of communication devices is relatively low, or it is easy to hire FE ports of other network providers with the reasonable cost.

- With the basic speed of FE connection from 128kb/s to 2Mb/s, the signal response speed of the IEC 60870-5-104 protocol is better than the IEC 60870-5-101 protocol. Moreover, it also supports 32-bit measurement (CP56Time2a).

Thus, the application of IEC 60870-5-104 communication protocol for SCADA systems of the distribution grid will basically overcome the limitations that the IEC 60870-5-101 were encountered. Based on TCP/IP network protocol, the IEC 60870-5-104 allows simple, low-cost communications, and easily exploits the telecommunications infrastructure of network service providers. However, the security in communication solutions must be specially prioritized when using public communication infrastructure [10-11].

#### IV. SIMULATION OF DAS ON THE DISTRIBUTION GRID

##### A. System diagram

In this paper, Feeder 471 and Feeder 476 of the 110kV Hoi An substation, presented in Figure 2, will be analyzed to simulate the DAS. The reason for choosing these feeders is the high penetration of photovoltaic rooftop systems around this area, which require timely supervision and control in order to ensure power quality and stability.

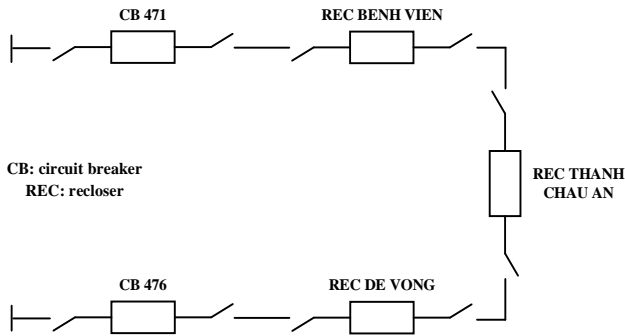


Fig. 2. Diagram of the simulation of the DAS tested on Feeder 471 and 476.

Initially, the communication signals of the control panels at Feeder 471, Feeder 476, Benh Vien, De Vong, Thanh Chau An and devices connected to the supervisory center are supposed to be normally operating, ensuring the backup power supply. Then, the incident cases will be assumed to evaluate the system. Possible incidents on Feeder 471 and Feeder 476 include: the incident on the segment from CB 471 to REC Benh Vien on Feeder 471 (case 1), the incident on the segment from REC Benh Vien to REC Thanh Chau An (case 2), the incident on the segment from CB 476 to REC De Vong on Feeder 476 (case 3) and the incident on the segment from REC De Vong to REC Thanh Chau An (case 4).

For example, the incident in case 1 is selected to analyze and build algorithm flowcharts, illustrated in Figure 3, to evaluate the system response. After identifying the incident location in the segment from CB 471 to REC Benh Vien, the protective relay will send the signal to switch off the CB at the beginning of Feeder 471. If the backup power source of Feeder 476 is available, the system confirms eligibility for the DAS program implementation. After that, the system will execute commands to switch off REC Benh Vien and switch on REC Thanh Chau An to continue supplying power for the segment that is not faulty. The automation of the distribution grid is

done by FDR and sectionalizers installed on the distribution power lines of the distribution grid combined with reclosing function (F79) of the CB equipped at the beginning of the feeder. In the system, the commonly used relays are: Instantaneous/AC Time Overcurrent (F50/F51), Neutral Instantaneous/Neutral Time Overcurrent (F50N/F51N), Reverse-Phase Overcurrent (F46). When an incident occurs on one of the two feeders (471 or 476), the system detects

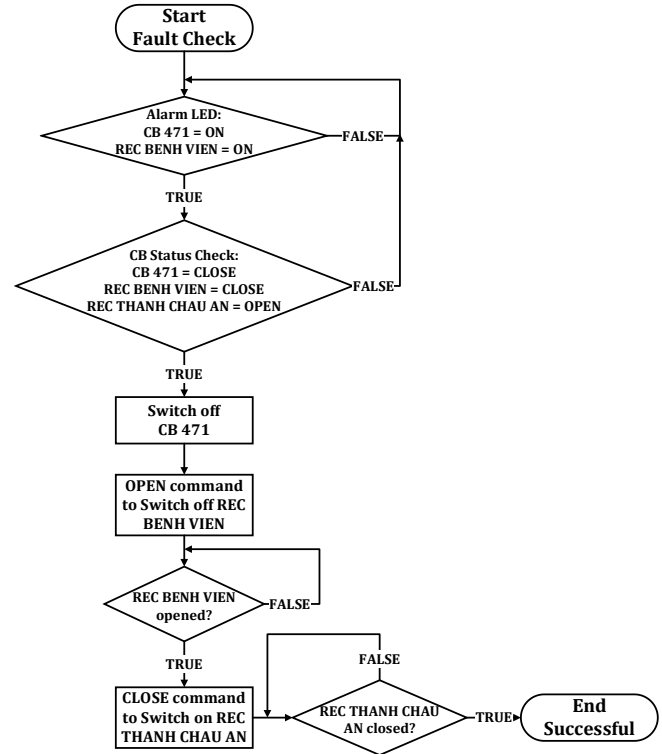


Fig. 3. Flowchart of problem-solving algorithms for incident case 1.

incidents thanks to information retrieved from the substation (breakers, relays, etc.), then execute commands to switch on/off CB to eliminate the incident and restore the normal operation of the grid. During this period, the faulty area is automatically isolated by existing devices on the two feeders of the distribution grid lines without the need of management and monitoring devices of the Regional Dispatch Center [11].

The protective time setting for overcurrent relay (standard inverse) in the program can be approximated by the following equation (based on the IEC 60255 standard):

$$t_{relay} = TMS \left( \frac{k}{\left( \frac{I}{I_s} \right)^\alpha - 1} \right) \quad (1)$$

where:  $t_{relay}$  is the tripping time (seconds),  $I$  is the fault secondary current (A),  $I_s$  is the relay pick-up current setting and  $TMS$  is the time multiplier setting. For standard inverse,  $k = 0.14$  and  $\alpha = 0.02$ .

LN	UC	UN	OA/VN	OB	OI	Description
QNE157_XT471	10	19	6511		QNE157_XT471	Bay local/remote-switch
QNE157_XT471	15				QNE157_XT471	Ext. interlocking (hardware)
QNE157_XT471	16				QNE157_XT471	Ext. interlocking (software)
QNE157_XT471	17				QNE157_XT471	Ext. interlocking command
QNE157_XT471	18				QNE157_XT471	Internal interlocking
QNE157_XT471	19				QNE157_XT471	Internal interlocking command
QNE157_XT471	20				QNE157_XT471	Bay blockings
QNE157_XT471	21				QNE157_XT471	Bay selection on monitor
QNE157_XT471	117				QNE157_XT471	Ext. interlocking command
QNE157_XT471	119				QNE157_XT471	Internal interlocking command
QNE157_XT471_LU	10	19	6231		QNE157_XT471	Line indicator
QNE157_XT471_LU	253				QNE157_XT471	Virtual switch for Topol. Col.
QNE157_XT471_LU	254				QNE157_XT471	Ext. ground ind. for Topol. Col.
QNE157_XT471_LU	255				QNE157_XT471	Infeed color for Topol. Col.
QNE157_XT471_LU	50000				QNE157_XT471	Voltage level
QNE157_XT471_471_76	10	19	6070		QNE157_XT471	471-76 Earth sw. position indication
QNE157_XT471_471_76	19	19	6071		QNE157_XT471	471-76 Earth sw. selection on monitor
QNE157_XT471_ALARM1	10	19	6111		QNE157_XT471	Alarm 1 F79 Disable
QNE157_XT471_ALARM1	11	19	6113		QNE157_XT471	Alarm 1 CB Spring Uncharged
QNE157_XT471_ALARM1	12	19	6115		QNE157_XT471	Alarm 1 F74 Fail
QNE157_XT471_ALARM1	13	19	6117		QNE157_XT471	Alarm 1 Communication Connection
QNE157_XT471_ALARM1	14	19	6119		QNE157_XT471	Alarm 1 AC OR DC MCB OFF
QNE157_XT471_ALARM1	18				QNE157_XT471	Alarm 1 Alarm indicator blockings
QNE157_XT471_ALARM1	19				QNE157_XT471	Alarm 1 Alarm indicator selected on monitor
QNR157_XT471_ALARM2	10	19	6121		QNE157_XT471	Alarm 2 Phase OC Instantaneous Trip 1
QNR157_XT471_ALARM2	11	19	6123		QNE157_XT471	Alarm 2 Reclose successful
QNR157_XT471_ALARM2	18				QNE157_XT471	Alarm 2 Alarm indicator blockings
QNR157_XT471_ALARM2	19				QNE157_XT471	Alarm 2 Alarm indicator selected on monitor
QNE157_XT471_ALARM3	10	19	6131		QNE157_XT471	Alarm 3 Phase OC Instantaneous Trip 2
QNE157_XT471_ALARM3	11	19	6133		QNE157_XT471	Alarm 3 Phase OC Instantaneous Trip 3
QNE157_XT471_ALARM3	12	19	6135		QNE157_XT471	Alarm 3 Ground O/C Trip 1

Fig. 4. Database of the entire system.

In addition, the automation of the distribution grid also comes with the functions of monitoring and remote control of sectionalizers. In order to fulfill this function, it is necessary to install RTUs and communication lines to receive information of sectionalizers located on the distribution lines. Based on the obtained information, the operators at the Dispatch Center will switch on/off the automatic CB to isolate the faulty element on the computer. At the supervisory center, it is necessary to install supercomputers to manage the operation of the distribution grid according to the geographical map and adjust the calculation automatically.

### B. Building database

In the SCADA system, every change of elements and devices in the system is monitored as updated data from time to time. The data constitutes the database system with the main purpose of communicating with devices connected to the system and updating information to the supervisory technician after receiving notification of changes in the system.

Dynamic information are called Variables or Tags, which are real-time units. The data is associated with the variables defined in the Variable List or Data List [12-13].

The order of database construction in SCADA systems is presented as follows [14-15]:

- Receiving update notifications: The SCADA system receives notifications that there is a change in the control objects.
- Querying the changes: The system displays the changes and system responses to the changes. For each affected object, information from the application model and the SCADA topology model for the device is retrieved. The application model system provides a to-do list (add, delete, modify) then apply to topology model.
- Updating topology model: The changes from devices and connections from the database to the SCADA topology model are compared, then the database to make changes validated are edited.

- Updating SCADA database: Updating device information to link status or field values.
- Operational control in testing mode.
- Providing online display of information and specifications for technicians.

Building database of the SCADA system is carried out thanks to the powerful support of MicroSCADA SYS600 9.4 Pro software provided by ABB Group.

To build the database, the "Object Navigator/Standard Function" tool is used by manually creating each data object or using Excel files to batch import all data of the substation into SYS600 software, as illustrated in Figure 4.

The software allows to initialize 8 signals corresponding to 8 addresses at a time. All processing objects are connected according to these following fields: Unit Number (UN), Object Address (OA), Object Bit Address (OB) [16].

### C. Developing program for the DAS

To build a control program for the DAS, it is first necessary to build a scheme of the two feeders 471 and 476 of 110kV substation. The "Display Builder" tool supports the construction of two feeders 471 and 476, which is illustrated in Figure 5. Based on the built-in database, the system diagram is designed using the "Actions/Objects Browser" tool.

The system control is simulated, using the control buttons: Fault simulation, Switching operations, Restore the system. The incident simulation program is implemented for the 4 cases mentioned in the previous section.

The SYS600 software uses the SCIL (Supervisory Control Implementation Language) programming language, illustrated in Figure 6, which is a high-level programming language specifically designed for the application of system monitor and control. All SYS600 application programs as well as most system configurations are built by SCIL [16-17].

After the implementation of the DAS, the response of the system is shown in Fig. 7.

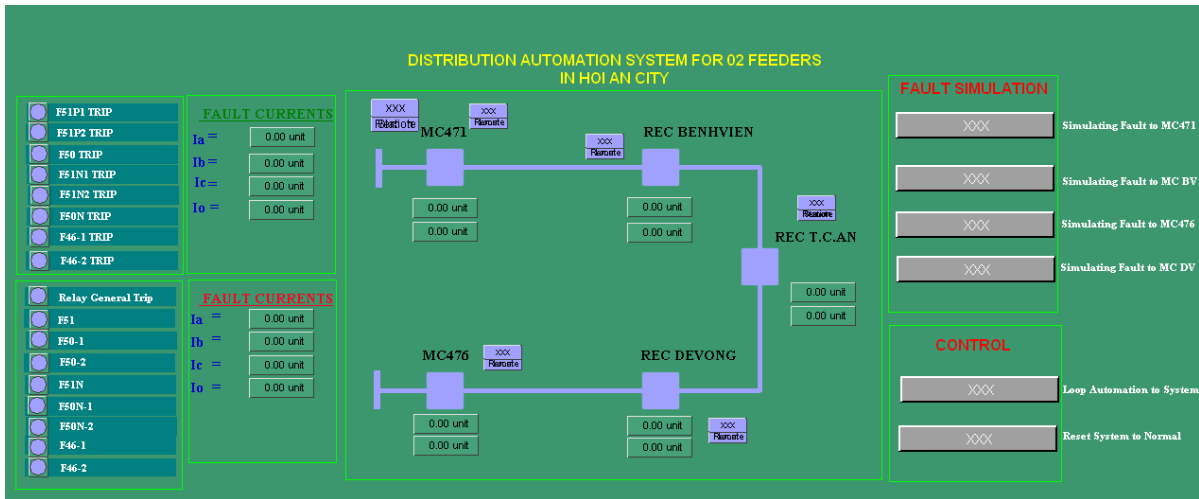


Fig. 5. Diagram of the DAS on Feeder 471 and Feeder 476.

```

[CODE SCIL]
=====
#local x_cot
x_cot=console_output("DAS thông báo:: Bạt đầu thực hiện LOOP AUTOMATION cho XT471E157 & XT476E157")
x_cot=console_output("DAS thông báo:: Step1: DAS kiểm tra xác nhận vị trí su co....")
#IF QNE157_XT471_CB471:P10==2 AND -
QNE157_XT471_ALARM3:P11==1 AND -
QNR007_CB:P10==1 -
#THEN #BLOCK
x_cot=console_output("DAS thông báo:: Step2: Xác nhận vị trí su co nam trong phân đoạn [MC471E157 và MC Benh Vien]!")
x_cot=console_output("DAS thông báo:: Step3: DAS kiểm tra nguồn du phong tu XT476E157 dam bao....")
#IF QNR005_CB:P10==2 AND -
QNE157_XT476_CB476:P10==1 -
#THEN #BLOCK
x_cot=console_output("DAS thông báo:: Step4: Xác nhận nguồn du phong tu XT476E157 dam bao!")
x_cot=console_output("DAS thông báo:: Step5: Xác nhận du điều kiện để thực hiện Loop Automation!")
x_cot=console_output("DAS thông báo:: Step6: DAS thực hiện lệnh [OPEN] [QNR007_MC Benh Vien]....")
#set QNR007_CB:pss10 = 1
#set QNR007_CB:P10 = 2
#IF QNR007_CB:P10 == 2 -
#THEN #BLOCK
x_cot=console_output("DAS thông báo:: Step7: Xác nhận đã [OPEN] [QNR007_MC Benh Vien] thành công!")
x_cot=console_output("DAS thông báo:: Step8: DAS thực hiện lệnh [CLOSE] [QNR005_MC Thanh Chau An]....")
#SET QNR005_CB:PSS10=1
#SET QNR005_CB:P10=1
#BLOCK_END
#ELSE x_cot=console_output("DAS thông báo: không [OPEN] được [QNR007_MC Benh Vien], DAS không thành công ! Ket thuc")
#IF QNR005_CB:P10==1 -
#THEN #BLOCK

```

Fig. 6. Part of control program for the DAS.

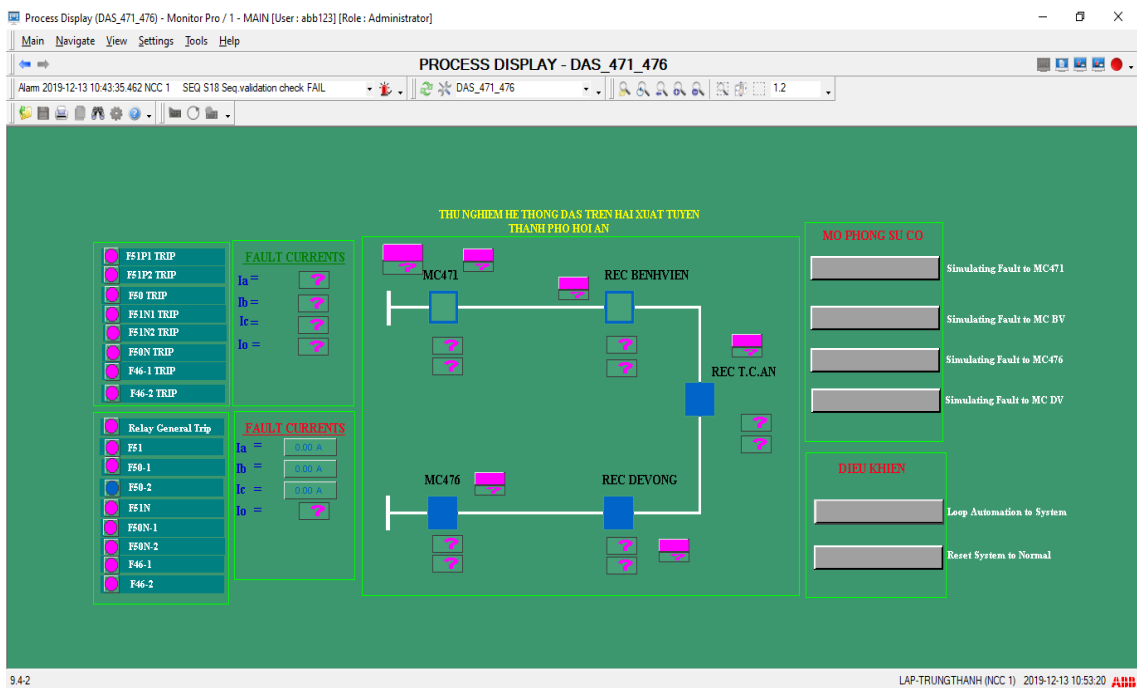


Fig. 7. The DAS performs automatic processing.

#### D. Assessment of the results

In the simulation, the system shows the ability to handle incidents quickly, accurately, independently and automatically, ensuring a safe isolation in different cases. The DAS contributes to solve power problems by limiting the area affected by power outages during power line incidents. Therefore, the application of DAS in the Vietnamese distribution grid in the future is necessary for the operation and management of the power system. In the near future, the proposed program will be tested and expected to put into operation. In addition, two indices System Average Interruption Duration Index (SAIDI) and System Average Interruption Frequency Index (SAIFI), which are used for the power system reliability assessment, must be considered in the upcoming research of the DAS. SAIDI and SAIFI are calculated based on (2) and (3).

$$SAIDI = \frac{\sum U_i N_i}{N_T} \quad (2)$$

$$SAIFI = \frac{\sum \lambda_i N_i}{N_i} \quad (3)$$

where:  $N_i$  is the number of power consumers and  $U_i$  is the annual outage time for location  $i$ ,  $N_T$  is the total number of consumers and  $\lambda_i$  is the failure rate.

Thanks to the application of the DAS, the power system reliability might be significantly improved, with SAIDI and SAIFI indices are expected to reduced by 20-25%.

#### V. CONCLUSION

In this paper, the authors have built the corresponding algorithm for the incidents that can occur in the Distribution Automatic System and have presented the application of MicroSCADA SYS600 9.4 Pro software in the construction of an actual DAS. The simulation case studies show that the DAS must perform the process of checking the signals of the devices to ensure stable connections before implementing the program. Therefore, the communication channel connecting the devices and the SCADA Supervisory center plays a very important role in the accuracy and safety of the DAS. Moreover, the system should use fiber-optic cable and have backup lines, network port monitoring equipment. Communication protocols must be selected consistently according to the National standards. The DAS must be programmed carefully, safely as well as calculated based on the conditions of various constraints: short-circuit current, number of switches of CB, types of short-circuit incidents (transient, steady state, etc.).

Lastly, the SCADA system at the Supervisory center should be evaluated and verified periodically, especially for communication devices connected to the system. All incidents due to SCADA software failures will lead to the interruption of the DAS program and the whole system. Thus, the study of

automation technology in the substation and the DAS aims to thoroughly implement the automation of distribution grid operation, bringing high efficiency in the operation of the Vietnamese power system in the future.

#### ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.02-2020.07.

#### REFERENCES

- [1] A. Daneels, W. Salter, "What is SCADA?", *International Conference on Accelerator and Large Experimental Physics Control Systems*, pp. 339-343, 1999.
- [2] D. J. Gauthier and W. R. Block, "SCADA communication techniques and standards," in *IEEE Computer Applications in Power*, vol. 6, no. 3, pp. 45-50, July 1993.
- [3] Gao, J., Liu, J., Rajan, B., Nori, R., Fu, B., Xiao, Y., Liang, W. and Philip Chen, C.L., "SCADA communication and security issues", *Security and Communication Networks*, 7(1), pp.175-194, 2014.
- [4] Lin, Chih-Yuan, and Simin Nadjm-Tehrani. "Understanding IEC-60870-5-104 traffic patterns in SCADA networks." In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security*, pp. 51-60. ACM, 2018.
- [5] Lê Kim Hùng, Nguyễn Thành "Ứng dụng hệ thống tự động lưới phân phối (DAS) để giảm thời gian và phạm vi mất điện khi có sự cố vĩnh cửu của lưới điện phân phối miền Trung", *Hội nghị toàn quốc lần thứ VI về tự động hóa-VICA VI*, 2005.
- [6] Jung, Sang Shin, David Formby, Carson Day, and Raheem Beyah. "A first look at machine-to-machine power grid network traffic." In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 884-889. IEEE, 2014.
- [7] Seung-Jae Lee, Myongji University, Korea, "Distribution Automation System (DAS)".
- [8] Goldenberg, N. and Wool, A., "Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems", *International Journal of Critical Infrastructure Protection*, 6(2), pp.63-75, 2013.
- [9] Kleinmann, A. and Wool, A., "Accurate Modeling of the Siemens S7 SCADA Protocol for Intrusion Detection and Digital Forensics", *Journal of Digital Forensics, Security and Law*, 9(2), p.4.
- [10] Teng, Le-tian, "The application of practical distribution automation technology in Shanghai distribution grid." In *2008 China International Conference on Electricity Distribution*, pp. 1-5. IEEE, 2008.
- [11] Scheidler, Alexander, L. Thurner, M. Kraiczy, and Martin Braun, "Automated Grid Planning for Distribution Grids with Increasing PV Penetration." In *6th Int. Workshop on Integration of Solar Power into Power Systems*, Vienna, Austria, 2016.
- [12] Thomas, Mini S., Seema Arora, and Vinay Kumar Chandna. "Distribution automation leading to a smarter grid." In *ISGT2011-India*, pp. 211-216. IEEE, 2011.
- [13] Minh Quan Duong et al., "Automatic Tool for Transformer Operation Monitoring in Smartgrid", *2019 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pp. 1-6, 2019.
- [14] Zhou Xue-song, Cui Li-qiang, Ma You-jie, "Research on Smart Grid Technology", *International Conference on Computer Application and System Modeling (ICCSM)*, vol. 3, pp. V3-599, 2010.
- [15] Kay N. Clinard, Raleigh NC, "Distribution Automation: Research and the Emergence of Reality", *IEEE Transactions on Power Apparatus and Systems*, vol. 103, no. 8, pp. 2071-2075, 1984.
- [16] MicroSCADA Pro SYS600 Application Design. ABB Group [Online], available at: <https://www.scribd.com/document/291072707/SYS600-Application-Design>
- [17] P281 MicroSCADA Pro SYS600 Operation. ABB Group [Online], available at: <https://www.scribd.com/document/309866157/P281-En-SYS600-94-Operation-14042015>



# A Study on Urban Traffic Congestion Using Simulation Approach

Phan Thi Kim Phung

Department of Industrial Management  
Can Tho University  
Can Tho City, Vietnam  
ptkphung99@gmail.com

Nguyen Truong Thi

Department of Industrial Management  
Can Tho University  
Can Tho City, Vietnam  
ntthi@ctu.edu.vn

Vo Thi Kim Cuc

Department of Industrial Management  
Can Tho University  
Can Tho City, Vietnam  
vtkcuc0012@gmail.com

**Abstract**—Urban traffic congestion has caused serious issues, such as environmental pollution, stress to drivers, and reduced safety levels despite the ongoing efforts and initiatives of urban planners to improve road infrastructure networks for long - term development. Focusing on temporary control strategies, this study aims to reduce traffic congestion at intersections through the consideration of traffic signal timing, two - stage left turn box, and traffic direction assignment. A simulation - based optimization model for traffic problems is developed to estimate vehicle travel time and queue length, which are important components of intersection performance. The traffic data of the main intersection located in Can Tho city, Vietnam was collected during peak hours to analyze and evaluate the existing and future traffic conditions. Our findings highlight the need for temporary control strategies to improve recurrent congestion and traffic - related air pollution exposure. Furthermore, the model in this study can be applied to investigate various traffic scenarios before implementing them in reality.

**Keywords**—simulation model; traffic management; traffic signal control; urban traffic congestion

## I. INTRODUCTION

Urban productivity is highly dependent on the efficiency of transportation systems to move people and freight between multiple departures and destinations. However, notable traffic problems, such as traffic congestion, road accidents, and environmental impacts, are severely spreading day by day with the rapid development of transportation systems around the world. Especially, in the urban areas of developing countries like Vietnam, the high rate of population results in a rapid increase in travel demand. The number of vehicles seems to be overwhelming as people's incomes increase while their prices decrease. In addition, the annual rate of increase in motor vehicles is growing more than that of the population. However, transportation infrastructure networks and law enforcement are insufficient to catch up with that growth, and therefore cause traffic congestion. Vietnam is considered as one of the leading countries with a high rate of traffic accident mortality and ambient air pollution due to transportation activities. Road traffic accidents caused an estimate of approximately 8,500 deaths recorded in 2016, according to the report of the United Nations [1, 2]. Furthermore, CO<sub>2</sub> is the primary source of greenhouse gas emissions (GHGs) emitted mainly from traffic through gasoline combustion and fuel evaporation. These emissions endanger public health, welfare, and contribute to global warming.

Traffic congestion at intersections has become an urgent problem for urban planners since intersections are considered as one of the most hazardous locations causing severe traffic accidents and heavy vehicle emissions for developing countries. However, it can be avoided or reduced with more effective use of traffic - management systems. The control strategies related to congestion reduction can be thought of as falling into two categories: permanent and temporary strategies. For permanent control strategies, urban planners can launch many mega transportation projects including road capacity expansion, building flyovers, tunnels, and public transport systems. It can be seen that such projects require huge building cost, land use, and time - consuming. The effectiveness and externalities of these expensive infrastructure investments have been questioned among urban, transport, and environmental planning scholars [3]. Some advocate that these projects generally conflict with environmental aims and architecture of the city, and would not yield expected results. While recognizing these practical limitations, more and more research is being carried out on the basis of urban traffic control systems. Some temporary control strategies, such as traffic signal timing, two - stage left turn box, and traffic direction assignment can be more effective strategies to manage and control traffic flows at intersections without much capital investment.

Computer simulation is a powerful tool to understand and analyze congestion problems since it can help decision - makers identify different possible options based on an enormous amount of dynamic data [3]. It is noted that the design and layout of signalized intersections can have influences on the efficiency of traffic movement. In this study, the temporary control strategies as aforementioned are simulated to estimate performance measures including vehicle travel time, queue length, and CO<sub>2</sub> emissions. Especially, environmental impacts are estimated based on the amount of emissions produced by different vehicle types on roads. A case study is conducted at the busiest intersection located in Can Tho city to determine the optimal answer for recurrent congestion. This city is the biggest city located in the Mekong Delta region, Vietnam with the rapid growth in the use of motorcycles as traffic tools.

The remainder of the study is organized as follows: Related work is introduced in Section 2. Then, a case study is developed in Section 3. Results and discussion of the study are presented in Section 4. The study ends with the conclusion in Section 5.

## II. LITERATURE REVIEW

Urbanization, known as the increasing number of the urban population, is one of the essential parts of national economic growth worldwide. It is estimated that over 50% of the population in the world are now living in towns and cities, in particular in Asia, Africa, and Latin America [2]. Despite the fact that it has a significant contribution to the GDP growth rate and provides opportunities for innovation and cultural development, urbanization may cause serious problems, such as air pollution, crime, traffic congestion, etc. Among them, the urban traffic congestion problem is one of the biggest challenges facing the local government authorities. Congestion can be classified as recurrent and non-recurrent. Recurrent congestion occurs when travel demand exceeds road capacity, while non-recurrent is original from disabled vehicles, construction zones, adverse weather, and special events [4]. Both types of congestion result in rising transportation costs and reducing safety levels for road users. Besides, motor vehicle emissions are the major source of releasing large amounts of pollutants to the environment as there are NO<sub>x</sub>, HC, CO, CO<sub>2</sub>, and PM 2.5. These toxic air pollutants can cause adverse health effects for people that work and live in most metropolitan areas, especially during peak hours. Therefore, many efforts and policies have been done to resolve traffic congestion over the past few decades. Investment in transport infrastructure and road capacity expansion is considered as the permanent control strategies that can improve traffic congestion. In this regard, Beaudoin et al. (2015) state that using public transport systems, such as city buses, trolleybuses, passenger trains, and ferries can reduce pressure on traffic congestion and air pollution [5]. In addition, through in-depth analysis, the study of Tennøy et al. (2019) highlights that road capacity expansion can handle future traffic growth in small and large cities of Norway [6]. However, this is the subject of debate. Litman (2013) states that it is only effective in the short-run, but would decline in the long-run [7].

Congestion problems at signalized intersections are highly related to traffic signal timing and traffic density [8]. Researchers and urban planners thus have shown great interest in traffic signal control systems to improve traffic efficiency and reduce traffic emissions as environmental issues have received public concern. The traffic signal optimization and signal settings based on delay minimization were firstly introduced in 1958. Following this interest, numerous further research studies on traffic congestion have been investigated for optimizing signal timings. Brian Park et al. (2009) in this regard develop a stochastic signal optimization method to design the signal timings when there are fluctuations in traffic demands. The consideration of fuel consumption and vehicle emissions in conjunction with traffic signal optimization is also mentioned in this study [9]. Yuan et al. (2014) identify traffic bottle roads in urban transportation networks with two factors including signal timing at intersections and properties of left-turn lanes and straight-through lanes of roads. A model is presented to reduce the impacts of traffic congestion arising from bottlenecks by optimizing signal timing at intersections [8]. Kou et al. (2018) establish a multi-objective model to minimize delay, stops, and vehicle emissions. [10]. Furthermore, Villagra et al. (2020) focus on the optimization of traffic light cycles to improve vehicle flows without any additional costs [11]. Beside numerous research studies on traffic signal timing, there is quite limited research on traffic direction assignment and a two-stage left turn box despite

their effectiveness to enable smooth traffic flows at signalized intersections. Zhao et al. (2013) propose exit lanes for left-turn (EFL) control via a mixed-integer nonlinear program with a combination of geometric layout, main signal timing, and pre-signal timing. The resulting study shows that EFL control leads to an increase in the capacity of intersection and to reduce average vehicle delay and queue length [12]. The use of a two-stage turn box at the signalized intersection as an efficient way to reduce the congestion and turning conflicts between vehicles can be found in the study of Ohlms and Kweon (2018). They propose two bike boxes and two-stage left turn boxes in the way that facilitate bicycle travel through intersections. The case study was conducted in Virginia [13].

There are various uncertainty factors affecting congestion problems such as the number of vehicles, travel time, and vehicle speed. Simulation optimization is widely used by urban planners and researchers due to its advantages over conventional methods in the consideration of uncertainty factors. Melouk et al. (2011) propose a Tabu search-based simulation-optimization approach to evaluate and mitigate traffic congestion. The study shows that simulation optimization can strongly support urban planners and local governments to tackle complex transportation planning tasks and road designs [14]. J. Javid & Jahanbakhsh Javid (2018) propose a micro simulation to estimate travel time variability caused by traffic incidents. The study integrates real time data including geometry, weather, traffic speed, and an extensive incident inventory database collected from highways [4]. Discrete event simulation (DES) model is applied to analyze an urban traffic signal control and improve the traffic flows of intersection. Kamrani et al. (2014) develop a simulation model for the traffic of two adjacent T-junctions located in Johor city, Malaysia during peak hours [3]. Bakhsh (2020) proposes a DES model to reduce vehicle cycle and waiting time at a roundabout intersection, and then suggests alternative system designs to optimize vehicle flows [15].

Tackling traffic congestion is a major concern for low-income countries. In Vietnam, numerous research efforts have been done on the impact of driver behavior on traffic flows so far. Nguyen-Phuoc et al investigate the prevalence and factors associated with the use of turn signal at intersections in Da Nang city as turn signal is a major cause of traffic crashes [16]. Quan-Hoang et al explore factors associated with behavior intentions of urban development and new transit alternative in Ho Chi Minh city [17]. Therefore, the motivation of this study is to provide supports for eco-friendly design and development of the busiest signalized intersection located in Can Tho city, Vietnam. This 4-way intersection is a connecting point of many important sites of the city and is characterized by a large number of daily traffic movements, especially at peak hours. Furthermore, this city shows a fast-growing number of motorcycles since Can Tho people use motorcycles as their first daily travel mode choice. In 2019, there were approximately 827,079 personal motorcycles. This number is estimated to increase substantially over the coming years. Traffic congestion from motorcycles is placed on full alert. Therefore local government authorities devote much of their attention to improving the existing traffic congestion from an efficiency, safety, and environmental benefit standpoint. In this study, the local traffic congestion is identified its causes and suggests the temporary control strategies including traffic signal timing, two-stage left turn box, and traffic direction without incurring considerable expenses. The effectiveness of proposed control strategies is

modeled and simulated using Arena simulation software to estimate performance measures including vehicle travel time, queue length, and vehicle emissions.

### III. PROBLEM DESCRIPTION

#### A. Case study description

Can Tho is the biggest city in the Mekong Delta region, Vietnam with a population density of 858 people/km<sup>2</sup>, much higher than the average across the country of 290 people/km<sup>2</sup>. The city has 9 districts and 85 sub - districts. Ninh Kieu is the city centre with the number of population, accounting for about 22.7% of the total registered residential population. As a centralized district, millions of immigrants have arrived here for working and educational purposes. Therefore, the demand for transportation has grown at a very fast rate. A large part of transportation is made through densely populated areas, causing serious traffic congestion and air quality impact.

The city has struggled with excessive traffic congestion due to rapid urbanization and limited road infrastructure in recent years. As mentioned in the previous section, congestion problems often occur at intersections, and therefore are required special attention in the planning and designing of intersections. Among intersections in Ninh Kieu district, Mau Than - 3/2, as depicted in Fig 1, is the most important element in the transportation networks of the city since it is located in a centralized traffic zone with many important sites of the city including universities, supermarkets, and hospitals. This 4 - way intersection is the meeting place for vehicles that are traveling from and to 4 major roads with high traffic density. Direction 4 is a three - traffic lane road on each side, while the other directions are two - traffic lane roads. Traffic volumes vary by the direction of travel, time of day, and day of the week. From a month's observation at the intersection, we found that directions 1 and 4 have a high concentration of traffic, comprising 58,01% of total vehicles approaching the intersection. Despite the presence of traffic police to handle road violations and prosecute traffic offences in the peak hour period (4:30 PM – 6:30 PM), substantial queues of vehicles in directions 1 and 3 normally create a longer waiting time.

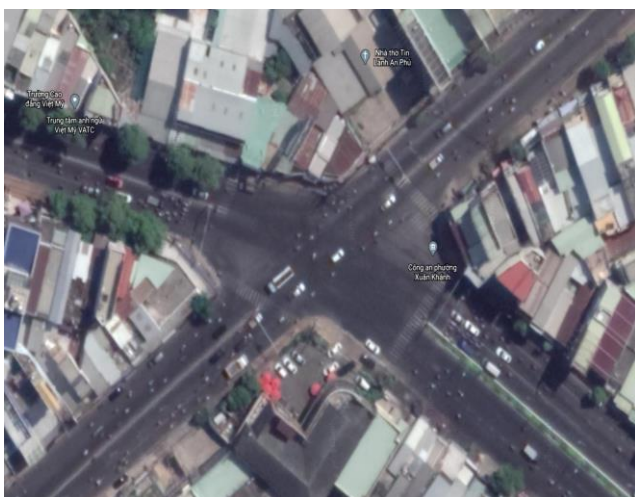


Fig. 1. General layout of Mau Than – 3/2 intersection in the road network of Can Tho city

As given in Table I and Fig 2, the average traffic volumes in the peak hour period are about 58,636 motor vehicles. The highest daily traffic volumes of all directions are found on

Monday and the most congested time of day is between 5:00 PM and 6:00 PM. The statistic shows that motorcycles are the dominant choice of road transport, comprising more than 94% of vehicles approaching the intersection. Research conducted by Nguyen-Phuoc et al. show that this average rate is about 78% in Da Nang, Vietnam [16]. Left – turn, right – turn, and straight through movements from such vehicles normally result in congestion problems, while the inefficiency of existing infrastructure and road capacity cannot hold the heavy traffic loads. Especially, left – turn movements, accounting for about 20.61%, are the high - risk movements and have a greater potential for conflict with opposing traffic flows. It is estimated that the trip in the peak hours can take a vehicle more than 39 seconds longer than the same trip in the off – peak hour period (8:00 AM – 10:00 AM). Furthermore, the concentration of a large number of vehicles in this highly populated area can generate an emission problem that seriously affects the quality of the urban environment and quality of life. Long - term exposure to traffic pollution or short - term exposure at higher pollution levels can cause significant health risks, especially an increase in the rate of decline of lung function. By understanding the harmful pollutant emissions that come from internal combustion engines, effective strategies for congestion problems must be developed and implemented shortly to protect the city's environment, and to strengthen the city's resilience to climate change.

TABLE I. AVERAGE NUMBER OF VEHICLE TYPES BETWEEN DIRECTIONS

Direction	Vehicle types			Total
	Motorcycles	Cars	Trucks	
D.1	15,431	697	276	16,404
D.2	12,029	623	106	12,758
D.3	11,381	531	113	12,025
D.4	16,596	655	198	17,449
<b>Total</b>	<b>55,437</b>	<b>2,506</b>	<b>693</b>	<b>58,636</b>

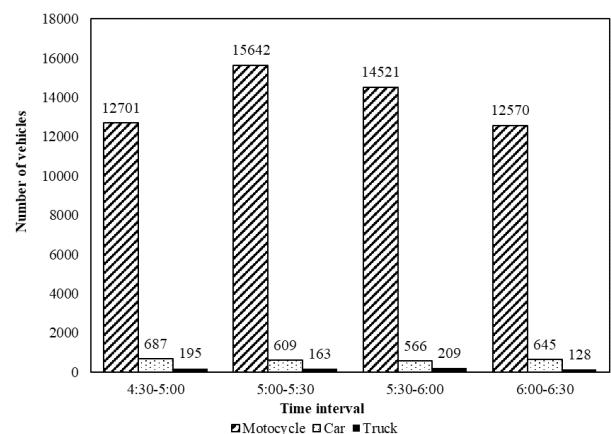


Fig. 2. Average number of vehicles on roads during the peak hour periods (4:30 PM – 6:30 PM)

The proposal for road capacity expansion and building flyover projects is introduced by city planners in response to the presence of such congestion problems. However, it is a debated project proposal for high building costs and has an enormous deforming effect on the form of the city. Even with

the priority to be given to project investment, it also raises a question about how heavy traffic flows can be improved. Therefore, the city is looking for more comprehensive and successful urban and transport planning solutions to keep the current congestion under control. In this study, basic traffic – based actions, including two - stage left turn box, traffic direction assignment, and traffic signal timing, are evaluated in terms of performance measures such as vehicle travel time, queue length, and vehicle emissions. Existing traffic data input such as road distances, number of motor vehicles on roads, time interval between vehicles, and vehicle speed at the intersection was collected by digital videos to establish a base point for assessing the traffic impacts. It is noted that this data was not collected during inclement weather or unusual traffic conditions such as roadway maintenance. Their possible impacts of each control strategy on traffic flows are discussed in the next part.

#### B. Logic model formulation

Following the case study, the logic model is constructed based on the three segments including existing traffic flows from directions, traffic control strategies, future traffic flows from directions, as represented in Fig 3.

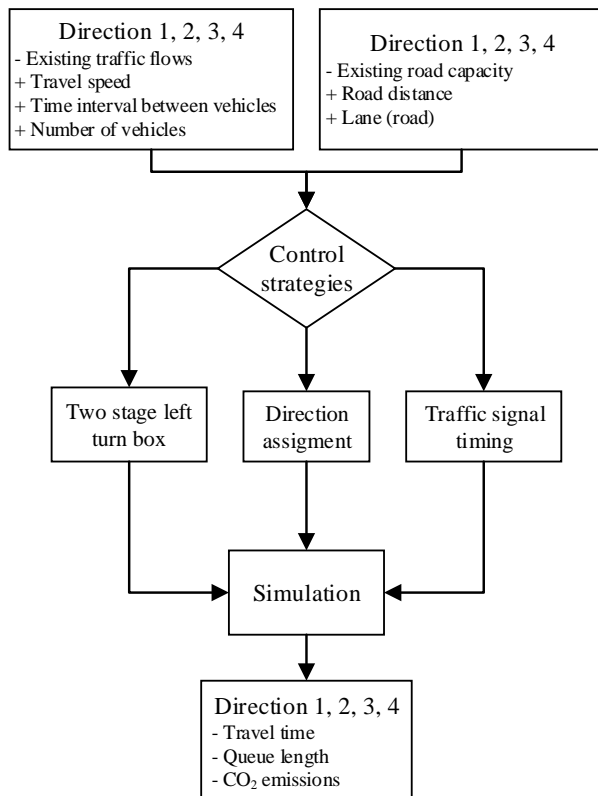


Fig. 3. Logic model based on the case study

In this study, the travel distance for a vehicle from a congested point to an exit point of each direction is assumed to be 300 meters. Traffic data such as time interval between vehicles and number of vehicles is collected at these points. The average vehicle travel time and the queue length for that distance are used to evaluate the traffic congestion states at the intersection.

#### IV. RESULTS AND DISCUSSION

This section shows the simulation results obtained from the case study. As mentioned earlier, our focus is to provide

the local government the temporary decisions to avoid or mitigate the negative effect of slow traffic flows on road users. In this study, computer simulation is a useful tool to decide which solutions would be appropriate for the existing traffic congestion at the intersection. The traffic problem models are constructed and solved using ARENA simulation software, Version 14.0 to represent the existing and future traffic states during the peak hour periods. These models are simulated for 120 minutes and 100 replications. It is noted that each scenario represents each individual decision to be made.

#### A. Existing traffic state (Scenario 1 – S1)

During a simulation run, the number of vehicles entering the system is 59,239, while the output is about 58,415. Fig 4 illustrates the distribution of number of motorcycles, cars, and trucks per second. The longest queue is found in the interval of 5:10 PM to 6:05 PM.

Validation is our concern when constructing the correct model, therefore the model of the existing traffic state is checked and verified by comparing its behavior with real - world conditions. Region R2 is selected to analyze since there is less variation in the average travel time. Note that the average travel time refers to the time required for a driver to finish 300m. The resulting value obtained from the model and the observed value obtained from data collection in Table II show that the model is valid when the 90% confidence interval for the percentage error of the model is below 10%. Therefore, the potential temporary strategies are then discussed and evaluated for further improvement of traffic congestion.

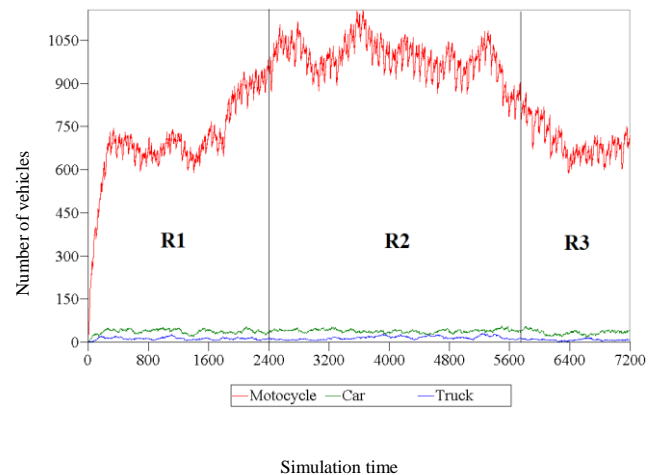


Fig. 4. Distrution of number of vehicles

TABLE II. AVERAGE TRAVEL TIME OF VEHICLE TYPES

	Motorcycle	Car	Truck
<b>Direction 1</b>			
Resulting value (seconds)	149	188	205
Observed value (seconds)	153	184	221
Percentage error (%)	2.6	2.2	7.2
<b>Direction 2</b>			
Resulting value (seconds)	120	138	158
Observed time (seconds)	115	131	175
Percentage error (%)	4.35	5.34	9.71
<b>Direction 3</b>			
Resulting value (seconds)	116	115	134
Observed time (seconds)	109	122	147
Percentage error (%)	6.42	5.74	8.84
<b>Direction 4</b>			
Resulting value (seconds)	80	88	89
Observed time (seconds)	85	97	98
Percentage error (%)	5.88	9.28	9.18



### B. Future traffic state (Traffic signal timing - Scenario 2 – S2)

Traffic signals are an essential aid tool to organize vehicles, to avoid accidents, and to improve the traffic flows. Traffic signal timings involve determining the sequence of operation and assigning how much time needed for red, yellow, and green indications. However, a wrong set of signal timings can significantly create vehicle queues, travel time, and a large amount of traffic emissions at the signalized intersection when green or red time is too long or short. Besides ensuring the safe crossing of traffic, traffic signals can help to reduce vehicle waiting times by appropriately adjusting cycle lengths. A cycle length is defined as the time in seconds required to display all signal indications before returning to the initial indication of the cycle. The existing signal cycle length of 80 seconds is not appropriate, and results in increased travel time and reduced safety levels. Therefore, optimization of traffic signal timing is important to minimize delays, stops, and vehicle emissions. Directions 1 and 3 get the red indication, then it is the turn for directions 2 and 4 to receive a green indication.

### C. Future traffic state (Traffic direction assignment - Scenario 3 – S3)

Although there are fewer conflicts between right-turning vehicles and opposing traffic right, such movements have also significant influences on traffic flows in the peak hours. According to our observation, the long vehicle queue length of direction 2 results from the limited road capacity to hold the large number of motorcycles approaching the intersection area. Meanwhile, the proportion of motorcyclists making a right turn at this direction (38.56%) is higher than the other directions. The existing traffic flows can be mitigated by providing motorcyclists an alternative roadway for right - turning. A “Right Turn Signal” sign is posted at direction 5 to indicate that motorcyclists must make right - turn movements to proceed. Furthermore, the same approach can be used for directions 1 and 4 when all right - turning vehicles are allowed to make a turn at the signalized intersection.

### D. Future traffic state (Two - stage left turn box - Scenario 4 – S4)

Left - turn movements at the intersection are a major source contributing to much traffic congestion. According to our observation, 16.01% of drivers from direction 1 make a left turn, while 50.16% of drivers from direction 3 proceed straight through the intersection. The existence of such movements from both directions can easily aggravate massive travel time delays, queue lengths, and crash risks for road users. The design of a two - stage left turn box helps to reduce potential turning conflicts between vehicles within the intersection area. It is a designated area at the head of traffic lanes at the signalized intersection. When using this option, a motorcyclist moves forward and turns into the box, and then waits until the next green signal indication before proceeding. He/she may take about 46 seconds to finish a two - stage left turn box.

### E. Future traffic state (Combination of proposed strategies - Scenario 5 – S5)

The two - stage left turn box, traffic direction assignment, and traffic signal control strategies have their advantages in terms of safety and traffic levels. Therefore, these individual

strategies are integrated to search for a better scenario, as illustrated in Fig 5.

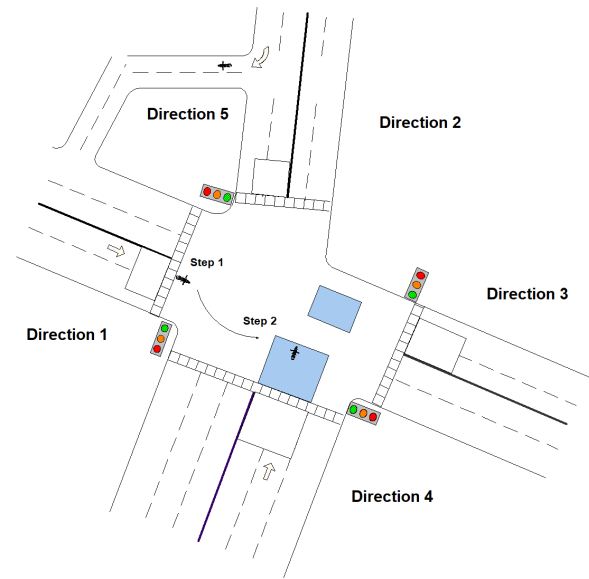


Fig. 5. General layout of Scenario 5

### F. Summary of simulation output

The selection of an appropriate scenario depends on how it influences the vehicle traffic flows. In this study, the performance evaluation is done in terms of the average vehicle travel time and queue length, which are investigated through the simulation results for 6 scenarios, as displayed in Fig 6. S1, S2, S3, S4, and S5 are represented for the strategies in the peak hour periods, while S6 is for the off - peak hour periods. It is seen that there is a strong relationship between travel time and queue length. A reduction in travel time may result in a shorter queue length. S6 is considered as a baseline scenario when there is no congestion. This scenario can be used to compare with the other alternative scenarios.

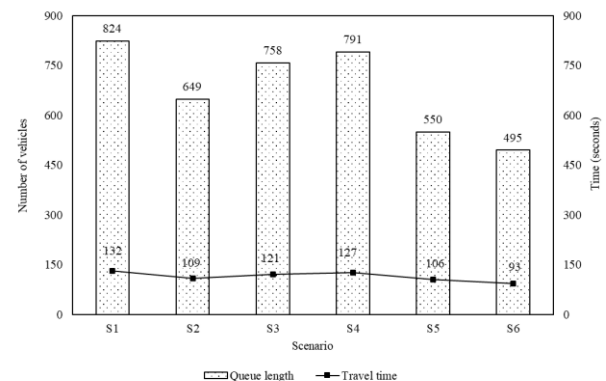


Fig. 6. Average queue length and travel time for scenarios

Many factors that affect vehicle emissions are driving conditions, age, vehicle type, and fuel type. In this study, the vehicle emission levels are simply estimated based on CO<sub>2</sub> emission rates of motorcycle, car, and truck on roadways. It is noted here that motorcycle emission rates of CO<sub>2</sub> are lower than those from cars and trucks, but a large amount of emissions at the intersection is produced by motorcycles since they are the dominant transport mode. Controlling emissions from not only motorcycles but also other vehicles would be



highly effective at decreasing Can Tho's air pollution. The effects of control strategies on total carbon emission levels of each vehicle type, as displayed in Fig 7, can provide opportunities to reduce environmental impacts. The results indicate that travel time has a strong correlated relationship with emission levels. An increase in travel time may result in a large amount of air pollution. Following this consideration, S1 achieves the highest amount of emissions as compared to the other scenarios, while the outcome with the greatest reduction of emissions to air is found in S6.

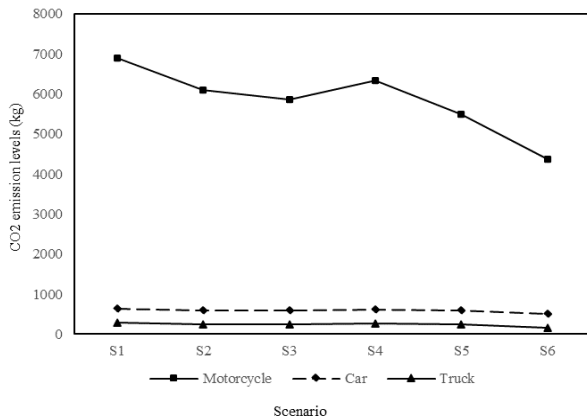


Fig. 7. Total CO<sub>2</sub> emission levels emitted from each vehicle type for scenarios

After accomplishing a series of simulation trials, the best scenario is determined by comparing the average travel time, queue length, and carbon emissions between scenarios. Scenario 1 (S1), representing the existing traffic conditions without using any proposed strategies, has the longest queue length and travel time as compared to other scenarios. Results indicate that there is a likely need to consider the proposed scenarios (S2, S3, S4, and S5) because they can help to reduce congestion at the intersection by eliminating bottlenecks and allowing traffic to flow more smoothly. Across these scenarios, the average travel time can be reduced by 2.04%. Although there is not much improvement in the traffic performance between S1 and S4 in this case, S4 can provide motorcyclists a safer way to make left turn movements on a multi-lane road. S5 shows the most improvement in terms of traffic flows and air quality impact, therefore it is further analyzed to demonstrate its effectiveness.

In further analysis, factors affecting the amount of time a vehicle crosses the intersection are investigated through the simulation results obtained from S5. The findings show that signal cycle length is significantly associated with vehicle travel time, queue length, and traffic emissions. From the resulting Table III, it is suggested that the signal timing should be varied based on the traffic volumes of each direction between vehicles as they approach the intersection at each time interval. Direction 4 has two signal phasing schemes. One is designed for straight-through movements, while the other is for left turn movements. The use of left turn phasing gives drivers turning left an opportunity to make their turn before others are allowed to proceed on their ways, therefore it is designed to start 10 seconds earlier than straight-through phasing. With an optimum cycle length of 86 seconds for all directions, vehicle travel time can be reduced by 10.92%. In this case, vehicles move faster on roads with a shorter delay time of 26 seconds. A driver may take about 1.8 minutes to

cross the intersection. Therefore, the length of the delays from the directions 1 and 4 become shorter.

TABLE III. TRAFFIC SIGNAL TIMING FOR SCENARIO 5

Traffic signal	Direction				
	D.1	D.2	D.3	D.4 (*)	D.4 (**)
Red light	44	44	44	44	53
Green light	40	40	40	40	31
Amber light	2	2	2	2	2
Cycle length	86	86	86	86	86
Travel time	131	98	112	84	

(\*) for straight-through movements

(\*\*) for left turn movements

## V. CONCLUSION

Traffic congestion and its consequences are gradually becoming an acute threat to the sustainability of cities. In order to build a city as a whole, the important decisions concerning the design of intersection are made in the transportation network planning process. Therefore, this study describes a simulation modeling approach to evaluate and compare the performance of different temporary control strategies for the busiest intersection. The proposed solution for reducing traffic congestion shows the feasibility and can be implemented with minimal capital investment.

In closing, the temporary control strategies can be applied to other intersections with some modifications depending on existing road networks. However, keeping congestion under control is an ongoing, never-ending task. These strategies should be reviewed every three to five years and more often if there are significant changes in traffic volumes or road conditions. Further study on traffic congestion in signalized intersections should be extended by considering different traffic-management policies.

## ACKNOWLEDGMENT

This work was supported by Can Tho University under Contract No. TSV2020-01.

## REFERENCES

- [1]. Road safety performance review. 2018, United Nations. p. 1-109.
- [2]. Stephens, C., Global Issues: Urban Health in Developing Countries, in International Encyclopedia of Public Health (Second Edition), S.R. Quah, Editor. 2017, Academic Press: Oxford. p. 282-291.
- [3]. Kamrani, M., S.M. Hashemi Esmail Abadi, and S. Rahimpour Golroudbary, Traffic simulation of two adjacent unsignalized T-junctions during rush hours using Arena software. Simulation Modelling Practice and Theory, 2014. 49: p. 167-179.
- [4]. J. Javid, R. and R. Jahanbakhsh Javid, A framework for travel time variability analysis using urban traffic incident data. IATSS Research, 2018. 42(1): p. 30-38.
- [5]. Beaudoin, J., Y.H. Farzin, and C.Y.C. Lin Lawell, Public transit investment and sustainable transportation: A review of studies of transit's impact on traffic congestion and air quality. Research in Transportation Economics, 2015. 52: p. 15-22.
- [6]. Tennøy, A., A. Tønnesen, and F. Gundersen, Effects of urban road capacity expansion – Experiences from two Norwegian cases. Transportation Research Part D: Transport and Environment, 2019. 69: p. 90-106.
- [7]. Litman, T., Transportation and Public Health. Annual Review of Public Health, 2013. 34(1): p. 217-233.
- [8]. Yuan, S., X. Zhao, and Y. An, Identification and optimization of traffic bottleneck with signal timing. Journal of Traffic and Transportation Engineering (English Edition), 2014. 1(5): p. 353-361.

- [9]. “Brian” Park, B., I. Yun, and K. Ahn, Stochastic Optimization for Sustainable Traffic Signal Control. *International Journal of Sustainable Transportation*, 2009. 3(4): p. 263-284.
- [10]. Kou, W., et al., Multiobjective optimization model of intersection signal timing considering emissions based on field data: A case study of Beijing. *Journal of the Air & Waste Management Association*, 2018. 68(8): p. 836-848.
- [11]. Villagra, A., E. Alba, and G. Luque, A better understanding on traffic light scheduling: New cellular GAs and new in-depth analysis of solutions. *Journal of Computational Science*, 2020. 41: p. 101085.
- [12]. Zhao, J., et al., Increasing the Capacity of Signalized Intersections with Dynamic Use of Exit Lanes for Left-Turn Traffic. 2013. 2355(1): p. 49-59.
- [13]. Ohlms, P.B. and Y.-J. Kweon, Facilitating bicycle travel using innovative intersection pavement markings. *Journal of Safety Research*, 2018. 67: p. 173-182.
- [14]. Melouk, S.H., et al., A simulation optimization-based decision support tool for mitigating traffic congestion. *Journal of the Operational Research Society*, 2011. 62(11): p. 1971-1982.
- [15]. Bakhsh, A., Traffic Simulation Modeling for Major Intersection (Sakarya University Journal of Science). *Sakarya University Journal of Science*, 2020. 24: p. 37-44.
- [16]. Nguyen-Phuoc, D.Q., et al., Turn signal use among motorcyclists and car drivers: The role of environmental characteristics, perceived risk, beliefs and lifestyle behaviours. *Accident Analysis & Prevention*, 2020. 144: p. 105611.
- [17]. Hoang, Q. and T. Okamura, Analyzing behavioral intentions in new residential developments of motorcycle dependent cities: The case of Ho Chi Minh City, Vietnam. *Case Studies on Transport Policy*, 2020. 8(1): p. 163-172.

# Optimizing Warehouse Storage Location Assignment Under Demand Uncertainty

Nguyen Truong Thi

Department of Industrial Management  
Can Tho University  
Can Tho City, Vietnam  
ntthi@ctu.edu.vn

Phan Thi Kim Phung

Department of Industrial Management  
Can Tho University  
Can Tho City, Vietnam  
ptkphung99@gmail.com

Tran Thi Tham

Department of Industrial Management  
Can Tho University  
Can Tho City, Vietnam  
tttham@ctu.edu.vn

**Abstract**—Assigning storage location in a warehouse has significant influences on space - consuming, storage capacity, and handling activities of goods. Therefore, in this study, storage management that integrates the determinant of production quantity is considered for a production system with multi - product and multi - period. A mathematical model is constructed to determine the optimal values including production quantity, inventory levels, and storage locations for goods with cost consideration. Besides that, a stochastic programming - based model is used to deal with the impact of demand uncertainty on warehouse performance over a planning period. The proposed models and solution approach are then investigated in a case study for the design of warehouse layout. Numerical results obtained from the model implementation and sensitivity analysis arrive at important practical insights. The resulting study shows that the study has a significant contribution to the effectiveness of warehouse layout.

**Keywords**—order picking, production planning, storage location assignment, demand uncertainty, warehouse layout

## I. INTRODUCTION

Logistics is widely known as the process of planning and executing the storage of goods and efficient transportation from the point of origin to the point of consumption. Logistics activities can be grouped into categories including packaging, order picking, inventory management, and product transportation. These activities have a close relationship with each other. Logistics decisions can highly impact on costs and bring countless benefits to a country's economy. Many studies have concluded that logistics costs usually contribute to a large share of the overall cost structure of a company. According to Singh, Chaudhary, & Saxena, warehousing cost in India is about 29% of the total logistics costs, whereas this cost in the USA is about 22% [1]. A warehouse is used for not only a place for storage but also an intermediate point for transportation and distribution of goods. Among warehouse activities, order picking is the most time - consuming and labor - intensive operation in logistics.

Over the past few years, many efforts of logistics companies have been made to improve order picking activities. One of ways to improve order picking operations is assigning storage location to appropriate goods so that handling activities of goods including loading, moving, and unloading by forklift trucks, stacker cranes, or conveying equipment can be easier and more effective. In addition, these activities impact costs and ability to respond to need. It is believed that a proper storage location is a part of solution to

achieve an optimization of operations that results in cost reduction and increased customer satisfaction. Despite the fact that storage assignment decisions have a direct effect on the frequency and distance travelled to pick up and retrieve goods, they rely heavily on the experiences and expertise of warehouse operators [2]. In addition, lack of information on production quantity and inventory levels between production managers and warehouse operators easily causes stock - outs and lost sales when occurring demand fluctuations. Under such circumstances, the decision - making process for storage location is usually a time - consuming and ineffective task.

Optimizing the use of storage locations can substantially reduce costs because storage and handling costs tend to increase geometrically as there is an increase in customer demand. In this study, production planning and storage location assignment are integrated in the decision - making process. Production planning involves the determination of production quantity, whereas storage location assignment allows assigning goods to proper storage locations after the production process. With this purpose, a joint optimization model is proposed in order to identify optimal storage locations and production quantities for a specific planning period, allowing minimizing production costs, handling costs, storage costs, and fixed costs throughout the planning period. Since time - varying demand patterns are increasing, the optimization problem for multi - product and multi - period under uncertain scenarios is suggested. The case study is conducted to demonstrate an example of practical location assignment problems with production and warehouse constraints.

The remainder of the study is organized as follows: Related works of production planning, order picking, and storage location assignment are introduced in Section 2. Then, problem description is provided in Section 3. Section 4 is basic assumptions and formulation of the model under study. Computational results and analysis of obtained results are presented in Section 5. The study ends with the conclusion in Section 6.

## II. LITERATURE REVIEW

Production planning, order - picking, and storage location assignment are fundamental decisions that impact the effectiveness and responsiveness of production and supply chain systems. Therefore, these decisions have attracted the attention of researchers over the past years. From the literature review, we found numerous studies on the use of mathematical optimization models to minimize total costs associated with production and storage operations.

### A. Aggregate production planning

Aggregate production planning (APP) has a critical role in companies due to its significant impact on production efficiency and cost - effectiveness. The decision - making problems for APP involve the determination of production quantity, inventory levels, available resources, and time needed to meet forecasted demand. According to the planning horizon, these decisions are classified as medium - term or tactical decisions that typically range from 3 to 18 months. The consideration of uncertainty on APP raises an important issue, which can be found in the study of Djordjevic et al and Jang & Chung. Djordjevic et al propose a new fuzzy model that considers time required to complete operations in the production and warehouse inventory [3]. Jang & Chung suggest a robust optimization approach to find optimal production capacity, workforce sizes, and inventory levels. This study considers unexpected variations in workforce levels from hiring and layoff uncertainty [4].

### B. Order picking systems

Order picking is defined as the process of retrieving items from specific storage locations to fulfill customer orders. This is a cost - consuming activity, accounting for about 35 - 50% of the total operating costs for a conventional warehouse [5]. Over the past decades, order picking has been an area of research interest with the purpose to minimize picking times and costs associated with handling activities. Two common order picking systems used in warehouses are parts - to - picker and picker - to - parts. Parts - to - picker is automated picking systems for storing and tracking goods by automatic storage and retrieval systems, whereas picker - to - parts systems require shippers to walk along picking aisles and pick up necessary goods from storage locations. Picker - to - parts systems are used more often than parts - to - picker because they require less capital investment as well as reduced operating costs. In the literature review provided by de Koster et al, more than 80% of enterprises in Western Europe use the picker - to - parts systems [6]. These systems are divided into low - level picking and high - level picking types. For low - level picking type, workers move along aisles to select goods from storage locations or racks. For high - level types, pickers use handling equipment such as forklift trucks or cranes to pick up goods and move to shipping stations. In this study, our focus is low - level picking types. Öncan & Çağırıcı consider the order batching problem with traversal and return routing policies. A mixed - integer linear programming (MILP) model is proposed to minimize total travel time for low - level picking systems employing human pickers [7].

### C. Storage location assignment

Storage location assignment is concerned with the determination of storage location for material and final products in warehouses. Many factors such as item properties, material handling systems, and customer demand have influences on the decisions related to storage allocation. Based on customer demand, warehouse operators normally make a decision about suitable storage locations where goods are placed and subsequently retrieved. According to the purpose of use, storage assignment policies can be classified as random

storage, dedicated storage, and class - based storage policies [5]. For the random storage policy, items can be randomly assigned to storage locations so that it can make good use of storage space utilization. Unlike this policy, the dedicated storage policy allows items to store in fixed storage locations. Meanwhile, the class - based storage policy integrates the random and dedicated storage policies, which group items into classes. Each class is determined by demand, type, and volume. The random storage policy is the simplest method, whereas the class - based storage policy is popular in use. In this study, our focus lies on the class - based storage policy. Manzini et al introduce a class - based storage assignment over a life cycle picking. MILP models are formulated to address technology selection, order picking systems, and storage location assignment [8]. Habibi Tostani et al introduce an integrated the class - based storage policy and dual shuttle cranes scheduling problem. Under this consideration, a bi - objective model considering multi - period planning and inventory is developed to minimize costs and energy consumption [9].

### D. Contributions

As aforementioned, the decisions related to production planning, order picking, and storage management have their advantages in terms of cost reduction and increased customer satisfaction. These decisions are highly affected by a high degree of uncertainty due to the dynamic nature of production systems and supply chain. In particular, unexpected demand fluctuations during a planning period make it hard for companies to satisfy customer demand. Therefore, deterministic solutions are not satisfactory and are even infeasible for real - world problems. Therefore, our motivation is to address the joint optimization of these decisions under demand uncertainty. In this respect, deterministic and stochastic programming models are developed and compared to determine the optimal production quantity and storage locations that minimize total costs including production costs, storage costs, and material handling costs. A typical warehouse is conducted to validate the proposed models.

## III. PROBLEM DESCRIPTION

ABC is a company that specializes in manufacturing a multi - product. The company is divided into three areas including production, warehouse, and output areas, as represented in Fig 1. Picking sequences start at the production area where items are manufactured and then packed into cardboard boxes of various sizes, depending on item type. These boxes are required to place on pallets, which in turn are put on forklift trucks to move to the warehouse as soon as finished production processes. The single floor warehouse under consideration acts as a regulating buffer between supplies and needs. It has a rectangular shape with a floor area of more than 382.61 square meters. This warehouse is responsible for the receiving, storing, and delivery of goods to maintain customer satisfaction. When receiving picking lists, picking operators use forklift trucks as handling equipment to retrieve goods from specified locations to the output area. This area is considered as the short - term storage for goods in transit without increasing inventory levels.

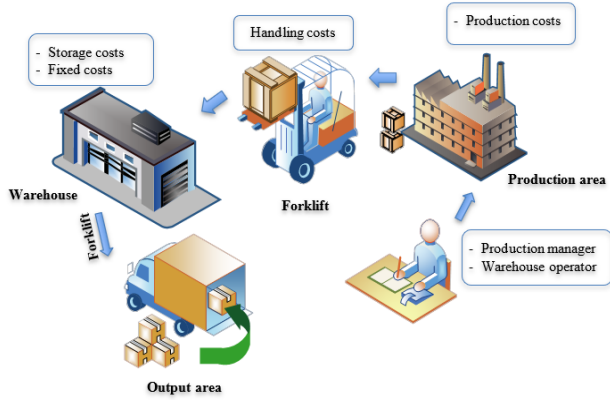


Fig. 1. Illustration of item flows in the company

In this study, the example of a pallet racking system in the warehouse is used as an illustration of the effect of storage location assignment on warehouse productivity. The system is designed with 5 pallet racks including A, B, C, D, and E that allow for the storage of palletized goods in horizontal rows with two levels. Each rack is able to hold a maximum number of 20 storage locations. Their storage space is limited to the number of items that can be stored. Rack A is located further to the production and output areas as compared to the others.

The warehouse has six parallel picking aisles between the pallet racks. Picking tasks can be done from both sides of the aisle without a significant change in position and also wide enough to allow a two - way travel. The movement of goods using forklift trucks in a warehouse can be horizontal and vertical. However, travel distances are often not optimized according to customer demand rates and the frequency of handling of goods, especially when there are changes in customer demand over time. Goods are located far away from point of origin to point of consumption, therefore causing longer picking time and increasing handling costs.

The company focuses on the production planning and storage location assignment for goods in the warehouse to keep high customer service levels while reducing production costs, storage costs, and handling costs simultaneously. It is noted that the production quantity at each period is varied based on the capacity of production lines, customer demand, and inventory levels.

#### IV. MODEL FORMULATION

The above case study is formulated as mixed - integer linear programming models (MILP), which aim at minimizing total costs including production costs, inventory costs, and transportation costs. Input data required for the model is customer demand, a number of items, and travel distances. Especially, travel distances between points via rectilinear aisles in this study are calculated based on rectilinear distances. The resulting optimization problem involves the determination of production quantity, storage locations, inventory levels, and other controllable variables for a specific planning period.

##### A. Assumptions

The mathematical models are developed based on the following basic assumptions:

- Each item is assigned to a single storage location at each period and cannot be replenished until this whole location is empty. However, stock - outs can be totally avoided for a planning period.

- Items are packed into boxes and placed on pallets after the production process at the production area. The size and shape of boxes handled are assumed to be fixed over a planning period.

- A forklift truck with a full payload of 2.5 tons is required to place loaded pallets onto racks for storage in the warehouse and retrieve when receiving customer orders.

- Costs associated with production planning and storage management are fixed over a planning period.

##### B. Indexes and parameters

We briefly review the mathematical formulation of the models. The indexes, parameters, and decision variables embedded in mixed - integer linear programming models are introduced in this section.

###### 1) Indexes

$p$ : Set of items ( $p = 1 \dots P$ )

$i$ : Set of storage locations ( $i = 1 \dots I$ )

$t$ : Set of periods ( $t = 1 \dots T$ )

$s$ : Set of scenarios ( $s = 1 \dots S$ )

###### 2) Parameters

$DW_i$ : Travel distance between production area and storage location  $i$  (m)

$DT_i$ : Travel distance between storage location  $i$  and output area (m)

$TT_p$ : Size of item  $p$  (storage space requirement of item  $p$ ) ( $m^3$ )

$TL_i$ : Size of storage location  $i$  ( $m^3$ )

$IDO_{ip}$ : Initial inventory level of item  $p$  at storage location  $i$

$CS_i$ : Cost of storage location  $i$  (VND)

$C$ : Travelling cost (VND/km)

$CI_{pt}$ : Unit inventory cost of carrying item  $p$  in period  $t$  (VND)

$CM_{pt}$ : Unit production cost of item  $p$  in period  $t$  (VND)

$ND_{pt}$ : Demand of item  $p$  in period  $t$

$ND_{pts}$ : Demand of item  $p$  in period  $t$  for scenario  $s$

$ML_p$ : Minimum production quantity of item  $p$

$MH_p$ : Maximum production quantity of item  $p$

$PR_s$ : Probability of scenario  $s$

$W$ : Loading capacity of forklift truck

$NR$ : Minimum inventory levels in period  $t$

$M$ : A big number



## 3) Decision variables

$NP_{pt}$  : Production quantity of item  $p$  in period  $t$

$NI_{pt}$  : Inventory level of item  $p$  in period  $t$

$NT1_{pit}$  : Number of times to transport item  $p$  from production area to storage location  $i$

$NT2_{pit}$  : Number of times to transport item  $p$  from storage location  $i$  to output area

$ID_{pit}$  : Inventory of item  $p$  at storage location  $i$  in period  $t$

$X_{pit}$  : Number of item  $p$  transported from production area to storage location  $i$  in period  $t$

$Y_{pit}$  : Number of item  $p$  transported from storage location  $i$  to output area in period  $t$

$Z_{pit}$  : 1, if item  $p$  is transported from production area to storage location  $i$  in period  $t$ ; 0

$V_{pit}$  : 1, if item  $p$  is transported from storage location  $i$  to output area in period  $t$ ; 0

$NP_{pts}$  : Number of item  $p$  produced in period  $t$  for scenario  $s$

$NI_{pts}$  : Number of item  $p$  kept in period  $t$  for scenario  $s$

$NT1_{pits}$  : Number of times to transport item  $p$  from production area to storage location  $i$  for scenario  $s$

$NT2_{pits}$  : Number of times to transport item  $p$  from storage location  $i$  to output area for scenario  $s$

$ID_{pits}$  : Inventory levels of item  $p$  at the storage location  $i$  at period  $t$  for scenario  $s$

$X_{pits}$  : Number of item  $p$  transported from production area to storage location  $i$  at period  $t$  for scenario  $s$

$Y_{pits}$  : Number of item  $p$  transported from storage location  $i$  to output area in period  $t$  for scenario  $s$

## 4) Deterministic model

When all parameters are known, the deterministic model can be formulated with the following objective function and constraints.

*a) Objective function:* The objective function of the model is formulated to minimize the total costs of production, storage occupation, and travelling for a planning period. These costs are proportional to the production quantity and stock - keeping units. Production costs reflect the expenses of producing products. Storage costs are the expenses associated with holding inventory in the warehouse. The economic impact of handling activities on the warehouse is especially considered in the study. Handling costs are determined based on the travel distances from the production area to storage locations, and then to the output area.

$$Z = \sum_p^P \sum_t^T CM_{pt} \times NP_{pt} + \sum_p^P \sum_i^I \sum_t^T CS_i \times V_{pit} + \sum_i^I \sum_t^T C \times DW_i \times NT1_i \times 2 + \sum_i^I \sum_t^T C \times DT_i \times NT2_i \times 2 + \sum_p^P \sum_t^T NI_{pt} \times CI_{pt}$$

## b) Constraints:

$$\sum_i^I X_{pit} = NP_{pt} \quad \forall p, t \quad (C1)$$

$$ML_p \leq NP_{pt} \leq MH_p \quad \forall p, t \quad (C2)$$

$$TT_p \times X_{pit} + IDO_{pit} \times TT_p \leq TL_i \quad \forall p, i; \forall t=1 \quad (C3)$$

$$TT_p \times X_{pit} + ID_{pi(t-1)} \times TT_p \leq TL_i \quad \forall p, i; \forall t \neq 1 \quad (C4)$$

$$\sum_i^I Y_{pit} = ND_{pt} \quad \forall p, t \quad (C5)$$

$$\sum_p^P V_{pit} \leq 1 \quad \forall i, t \quad (C6)$$

$$X_{pit} \leq Z_{pit} \times M \quad \forall p, i, t \quad (C7)$$

$$Y_{pit} \leq Z_{pit} \times M \quad \forall p, i, t \quad (C8)$$

$$Z_{pit} \geq V_{pit} \quad \forall p, i, t \quad (C9)$$

$$X_{pit} + IDO_{pi} - ID_{pit} = Y_{pit} \quad \forall p, i; \forall t=1 \quad (C10)$$

$$X_{pit} + ID_{pi(t-1)} - ID_{pit} = Y_{pit} \quad \forall p, i; \forall t \neq 1 \quad (C11)$$

$$\sum_i^I ID_{pit} = NI_{pt} \quad \forall p, t \quad (C12)$$

$$Z_{pit} = 1 \quad \forall p=r, m \leq i \leq n, t \quad r \in P; m, n \in I \quad (C13)$$

$$Z_{pit} = 0 \quad \forall p \neq r, m \leq i \leq n, t \quad r \in P; m, n \in I \quad (C14)$$

$$NT1_{pit} \geq \frac{X_{pit}}{W} \quad \forall p, i, t \quad (C15)$$

$$NT2_{pit} \geq \frac{Y_{pit}}{W} \quad \forall p, i, t \quad (C16)$$

$$NP_{pt}, NI_{pt}, NT1_{pit}, NT2_{pit}, ID_{pit}, X_{pit}, Y_{pit} \geq 0 \quad \forall p, i, t \quad (C17)$$

$$Z_{pit}, V_{pit} \in \{0, 1\} \quad \forall p, i, t \quad (C18)$$

Constraint (C1) sets the total number of items transported from the production area to the warehouse in each period. Constraint (2) ensures that the production capacity constraint of a plant cannot be violated. Constraints (C3) and (C4) limit the number of items that is assigned to a single storage location due to its storage space. Constraint (C5) states that customer demand of items in each period must be satisfied by a warehouse. Constraint (C6) ensures that each storage location can only keep one item type in each period. Constraints (C7), (C8), and (C9) assure that all items can only be assigned to a storage location once there exist item flows from the production area to the output area via that location. Constraints (C10) and (C11) state that changes in inventory levels between periods result from the difference between the inflow and outflow of items. Constraint (12) indicates that the minimum stock levels are required to avoid the occurrence of stock - outs. Some items must be stored in specific locations that facilitate a picking process. Therefore, constraints (C13) and (C14) specify special storage locations for these items. Constraints (C15) and (C16) represent the number of times transported from the production area to the warehouse and then to the output area. Constraint (C17) defines non -

negativity, and integer of decision variables. Constraint (C18) is binary constraints for all storage locations under consideration.

##### 5) Stochastic programming model

The proposed deterministic model assumes that all parameters including cost and demand are precisely known. As aforementioned, there is a need to look into uncertainty in the decision - making process. Demand uncertainty is a major source of risk. Therefore, the effects of uncertain demand on storage management are considered in this study. The proposed stochastic model takes into account the risks related to uncertainty in demand. Since scenario - based approach is successfully used in the previous work of Amin & Zhang [10], it is suggested to solve the proposed mathematical model under demand uncertainty. Following this method, uncertainty is represented by a set of different scenarios, which captures how the uncertainty might be in the future. Each scenario is associated with a probability level representing the expectation of the occurrence of a particular scenario. In order to formulate the stochastic programming model, new sets, parameters, and decision variables are added to the previous deterministic model.

##### a) Objective function

$$Z = \sum_p \sum_t \sum_s PR_s \times CM_{pt} \times NP_{pts} + \sum_p \sum_i \sum_t CS_i \times V_{pit} + \\ \sum_i \sum_t \sum_s PR_s \times C \times DW_i \times NT1_{is} \times 2 + \\ \sum_i \sum_t \sum_s PR_s \times C \times DT_i \times NT2_{is} \times 2 + \\ \sum_p \sum_t \sum_s PR_s \times NI_{pts} \times CI_{pt}$$

##### b) Constraints

$$\sum_i X_{pits} = NP_{pts} \quad \forall p, t, s \quad (C19)$$

$$ML_p \leq NP_{pts} \leq MH_p \quad \forall p, t, s \quad (C20)$$

$$TT_p \times X_{pits} + IDO_{pit} \times TT_p \leq TL_i \quad \forall p, i, s; \forall t=1 \quad (C21)$$

$$TT_p \times X_{pits} + ID_{pis(t-1)} \times TT_p \leq TL_i \quad \forall p, i, s; \forall t \neq 1 \quad (C22)$$

$$\sum_i Y_{pits} = ND_{pts} \quad \forall p, t, s \quad (C23)$$

$$X_{pits} \leq Z_{pit} \times M \quad \forall p, i, t, s \quad (C24)$$

$$Y_{pits} \leq Z_{pit} \times M \quad \forall p, i, t, s \quad (C25)$$

$$X_{pits} + IDO_{pi} - ID_{pits} = Y_{pits} \quad \forall p, i, s; \forall t=1 \quad (C26)$$

$$X_{pits} + ID_{pi(t-1)s} - ID_{pits} = Y_{pits} \quad \forall p, i, s; \forall t \neq 1 \quad (C27)$$

$$\sum_i ID_{pits} = NI_{pts} \quad \forall p, t, s \quad (C28)$$

$$NT1_{pits} \geq \frac{X_{pits}}{W} \quad \forall p, i, t, s \quad (C29)$$

$$NT2_{pits} \geq \frac{Y_{pits}}{W} \quad \forall p, i, t, s \quad (C30)$$

$$NI_{pts} \geq NR_p \quad \forall p, t, s \quad (C31)$$

$$NP_{pts}, NI_{pts}, NT1_{pits}, NT2_{pits}, ID_{pits}, X_{pits}, Y_{pits} \geq 0 \quad \forall p, i, t, s \quad (C32)$$

## V. COMPUTATIONAL RESULTS AND DISCUSSION

A good storage management can greatly increase productivity and storage capacity while reducing the costs of goods stored and picked. In this study, the deterministic and stochastic models are coded and solved using IBM ILOG CPLEX optimization studio software, version 12.4. The effectiveness of the proposed cost models is tested and performed under the consideration of 11 scenarios with different data sets of customer demand for the planning period. Each scenario is run on a Core i5 processor, 8GB RAM within 60 seconds, and reaches an optimality gap of less than 2%. The resulting scenarios can generate optimal production quantities, inventory levels, and storage locations for 10 item types in the three - period planning. The study is expected to assist production managers and warehouse operators in effectively identifying the picking sequence from the production area to the warehouse, and then to the output area.

The selected scenarios for analysis and discussion are listed in Table I. The changes in production, storage, and handling costs between scenarios result from the customer demand changes over the planning period. Each scenario has a probability distribution of demand. It is worth noting that the resulting deterministic model is represented by Scenario 1 to 10. Scenario 1 is considered as a baseline scenario with the probability level of 0.3, whereas Scenario 11 is the result of the stochastic model. Despite an increase in total costs as increasing customer demand, there is not much difference in the percentage of cost between scenarios. Among cost components, production costs make up a large share of total costs, ranging from 82.84% to 83.67%. The remaining percentage of costs are the costs associated with storage location - allocation in the warehouse. Noted that each selected location involves the fixed costs, so they should be optimized in such a way that the number of empty storage locations can be maximized. Following this consideration, the empty space of each location is also minimized. Handling costs, accounting for about 5%, are highly based on the travel distances from the production area to storage locations, and then to the output area. Travel distances impact not only on handling costs but also on picking time.

TABLE I. COST COMPONENTS OBTAINED FROM SCENARIOS

Unit: 1.000 VND

Scenarios	Probability	Production costs	Fixed costs	Storage costs	Handling costs
1	0.3	267.780	40.430	1.935	13.445
2	0.025	271.260	39.000	1.705	14.865
3	0.1	257.240	37.180	2.529	13.571
4	0.075	267.160	37.830	1.769	14.528
5	0.1	260.260	37.700	2.459	13.002
6	0.15	265.070	37.960	2.024	13.156
7	0.025	270.400	27.040	3.109	14.665
8	0.025	274.840	39.130	1.647	12.852
9	0.15	259.950	38.350	2.093	14.405
10	0.05	264.930	37.700	2.125	13.141
11		263.480	44.330	2.083	14.730

As aforementioned, variations such as consumption, delivery lead time, and destination need to be monitored closely. In this study, demand uncertainty is attached in the proposed stochastic programming model. This model can yield better results compared with the deterministic model in most scenarios in terms of total costs. This indicates that uncertain demand is a significant and fundamental element of a mathematical model. From the resulting Table I, the stochastic scenario 11 has a low - cost difference compared to the baseline scenario with a 0.05% change. A decrease in production costs of 1.28% would result in an increase in fixed costs of 0.9%. This is due to an increase in the number of storage locations which are not utilized to keep goods.

To show the effects of production costs on the objective function, the sensitivity analysis of deterministic and stochastic models is performed in Fig 2. The resulting analysis shows that the optimal storage location - allocation is very sensitive to changes in production costs. It can be seen that by increasing this parameter 10%, the values of objective functions are also increased. The deviations in total costs reveal that there is a need to look into forecasts to reduce risks in the decision - making process under demand uncertainty.

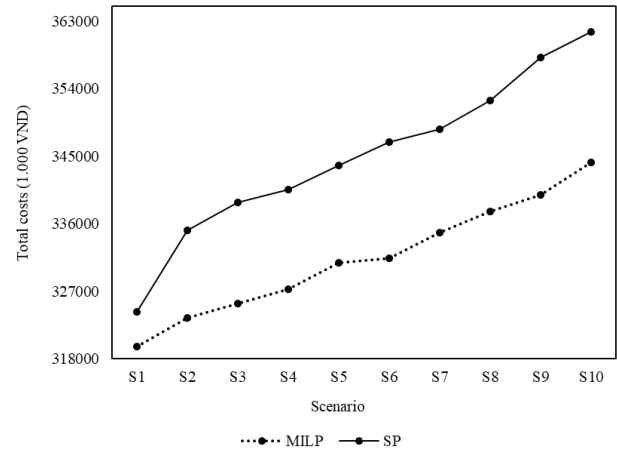


Fig. 2. Sensitive analysis of cost in deterministic and stochastic models

Storage location assignment is the main factor for transportation, as it determines the number of forklift trucks

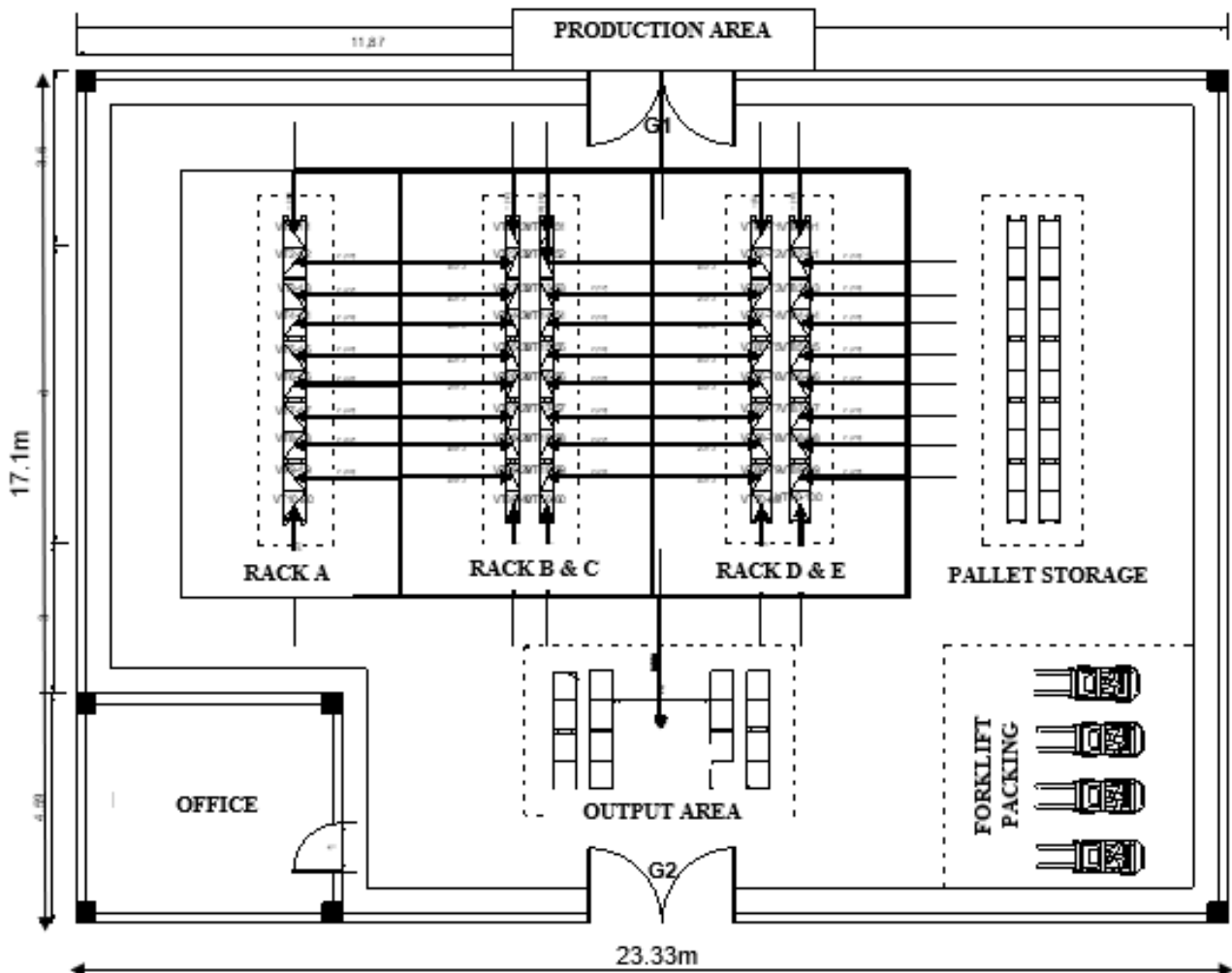
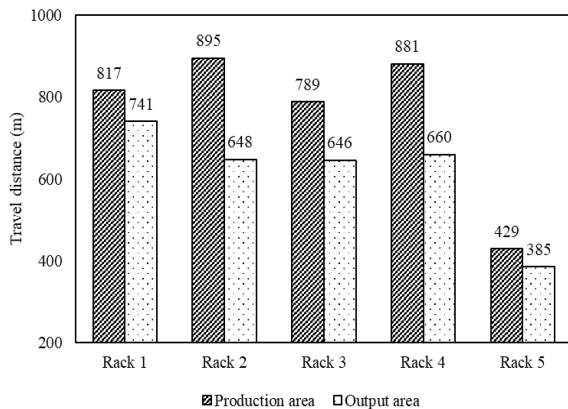


Fig. 3. Illustration of picking sequence in warehouse

or trips needed. Therefore, it should be designed in such a way that goods can be easily picked up and moved as needed. Therefore, optimizing the use of storage locations according to their space availability, inventory levels, and frequency of use is expected to shorten travel time for storing and retrieving items. Items with fast turnover in terms of consumption are suggested to store in locations closer to the input and output points. Moreover, by prioritizing fast - moving stock, the time to load and unload goods can be reduced significantly. In this section, the stochastic model (Scenario 11) is analyzed to see how high - demand rates should be kept in storage locations. The detailed layout of warehouse, as depicted in Fig 3, can permit easy stock handling and access to all stocks for inspection. A high picking density can be found in the storage locations of Racks C and D. This is due to their proximity to input and output points. Their storage space is always occupied by a large number of items in the planning period. As a result, pickers access the picking aisles of these locations more often than others. The picking volumes of these racks account for 52.05%. Some special storage locations of Racks A and B can only be available to keep items 1, 2, 3, and 4, whereas the others can be used for the remaining items. However, many unused storage locations can be found in Rack A due to limited demand. It is noted that all locations can be replenished at the beginning of the period so that there is no occurrence of stock - outs.

The resulting scenario 11 suggests that the optimal number of storage locations used for items is 73, whereas the warehouse has a maximum capacity of 100 storage locations. As a result, the total space utilization ratio of the warehouse is 73% after storage location - allocation. but there are only some of them with full utilization. The principle for the selection of storage locations is that their locations have to be centralized to achieve the shorter travel distances. As recorded in Fig 4, the total travel distances from the production area to the racks are further than that from the racks to the output area.



The number of products to be produced at the beginning of each period must be determined based on customer demand, production capacity, and storage capacity of the warehouse. From the resulting Table II, there are differences between customer demand and production quantity because of the influence of inventory. It is reasonable to allow safety stocks in response to fluctuations in customer demand. Following this consideration, minimum inventory levels of each period are established to reduce unnecessary loss and to satisfy customer satisfaction for the planning period. As a

result, the total inventory levels are 239 boxes, accounting for about 13.28% of production quantity.

TABLE II. PRODUCTION PLANNING FOR PRODUCTS IN THE PLANNING PERIOD

Items	Production quantity	Customer demand	Inventory levels
1	196	157	39
2	228	195	33
3	191	168	23
4	192	176	16
5	197	169	28
6	246	230	16
7	246	225	21
8	226	208	18
9	160	126	34
10	148	137	11
<b>Total</b>	<b>2,030</b>	<b>1,791</b>	<b>239</b>

## VI. CONCLUSION

Since the storage area in a warehouse is critical, effective storage management can facilitate effective storekeeping and comply with safety standards, and allow sufficient space for picking and delivery operations. This study presents the approach that production planning, order picking, and storage location are integrated in the decision - making process. Under demand uncertainty, the stochastic programming model for multi - product and multi - period is developed with the objective to minimize the total costs of production, storage, and handling activities. The resulting study shows that the model is able to apply for similar cases.

The consideration of 11 scenarios corresponding to customer demand with their probability is our limitation. In reality, due to the uncertain nature of the future, there is a need to provide more alternative scenarios to analyze. The small - scale problem under study can be solved optimally by using the exact optimization method with limited time. However, for large - scale problems, algorithms should be extended in further studies to generate appropriate solutions.

## REFERENCES

- [1]. Singh, R.K., N. Chaudhary, and N. Saxena, Selection of warehouse location for a global supply chain: A case study. *IIMB Management Review*, 2018. 30(4): p. 343-356.
- [2]. Hou, J.-L., Y.-J. Wu, and Y.-J. Yang, A model for storage arrangement and re-allocation for storage management operations. *International J. of Computer Integrated Manufacturing*, 2010. 23(4): p. 369-390.
- [3]. Djordjevic, I., D. Petrovic, and G. Stojic, A fuzzy linear programming model for aggregated production planning (APP) in the automotive industry. *Computers in Industry*, 2019. 110: p. 48-63.
- [4]. Jang, J. and B.D. Chung, Aggregate production planning considering implementation error: A robust optimization approach using bi-level particle swarm optimization. *Computers & Industrial Engineering*, 2020. 142: p. 106367.
- [5]. Wang, M., R.-Q. Zhang, and K. Fan, Improving order-picking operation through efficient storage location assignment: A new approach. *Computers & Industrial Engineering*, 2019. 139: p. 106186.
- [6]. de Koster, R., T. Le-Duc, and K.J. Roodbergen, Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, 2007. 182(2): p. 481-501.
- [7]. Öncan, T. and M. Çağırıcı, MILP Formulations for the Order Batching Problem in Low-Level Picker-to-Part Warehouse Systems. *IFAC Proceedings Volumes*, 2013. 46(9): p. 471-476.
- [8]. Manzini, R., et al., Modeling class-based storage assignment over life cycle picking patterns. *International Journal of Production Economics*, 2015. 170: p. 790-800.
- [9]. Habibi Tostani, H., et al., A Bi-Level Bi-Objective optimization model for the integrated storage classes and dual shuttle cranes scheduling in AS/RS with energy consumption, workload balance and time windows. *Journal of Cleaner Production*, 2020. 257: p. 120409.
- [10]. Amin, S.H. and G. Zhang, A multi-objective facility location model for closed-loop supply chain network under uncertain demand and return. *Applied Mathematical Modelling*, 2013. 37(6): p. 4165-4176.

# Optimization Design of a Compliant Tension Spring

Minh Phung Dang

Faculty of Mechanical Engineering, Ho  
Chi Minh University City of Technology  
and Education

Ho Chi Minh City, Vietnam  
phungdm@hcmute.edu.vn

Hieu Giang Le

Faculty of Mechanical Engineering, Ho  
Chi Minh University City of Technology  
and Education

Ho Chi Minh City, Vietnam  
gianglh@hcmute.edu.vn

Xuan Hoang Vo

Faculty of Mechanical Engineering,  
Industrial University of Ho Chi Minh City  
Ho Chi Minh City, Vietnam  
xuanhoangvo3994@gmail.com

Thanh-Phong Dao\*

Division of Computational Mechatronics, Institute for Computational Science, Ton Duc Thang University

Faculty of Electrical and Electronics Engineering, Ton Duc Thang University

Ho Chi Minh City, Vietnam  
daothanhphong@tdtu.edu.vn

**Abstract**—This study presents an optimization design for a compliant tension spring. Structure of the proposed spring is designed based on foldable mechanism to reach a large range of deformation. To improve the performances of the spring, a computational optimization process is implemented by an integration of the response surface method, finite element method, and multi-objective genetic algorithm. First of all, a 3D model of the spring is created. Then, design of experiments is built by central composite design. Simulations are performed to collect the numerical datasets. Subsequently, Kriging metamodel is used to approximate a relationship between the design parameters and the deformation. And then, the multi-objective genetic algorithm is utilized to search the Pareto-optimal set for the spring. The sensitivity of design parameters is analyzed. Finally, the predicted results are validated through finite element analysis.

**Keywords**—component, Compliant tension spring, Optimization, FEM, Kriging metamodel, Multi-Objective Genetic Algorithm Introduction

## I. INTRODUCTION

In the last two decades, compliant mechanisms have been received a great interest from academic researchers, industry, and practitioner [1, 2]. The fact is that compliant mechanisms have some excellent advantages such as a minimal components, high precision, free lubricant and friction in comparison with rigid-body counterparts [3–5]. Complaint mechanisms are playing a vital role in precise positioning engineering, robotics, and so on [6–8]. Especially in the field of assistive technology and rehabilitation system, compliant mechanism may be an excellent candidate to decrease the size of device and minimize the assemble counterparts. In the present study, a compliant tension spring is developed for use in the assistive technology and rehabilitation systems. Shape and structure of the spring are designed based on series of foldable leaf hinges so as to enlarge the displacement. However, design and analysis of the proposed spring has been facing challenges due to the spring is fabricated monolithically and there is a coupling between kinematic and mechanical behaviors. It induces a quite difficult in predicting and enhancing the performances of the spring. For a real application, the spring is desired to reach a good working strength and a wide range of stroke. Nevertheless, both these objectives are conflicted to each other. Until now, this problem is still hot topic for researchers in the field of compliant mechanism.

The goals of this article are to propose an approach to conduct a multiple-objective optimization problem for the compliant tension spring. Shape and structure of the spring are designed. And then, key geometrical factors and output performances are determined. Simulations are carried out to collect numerical datasets. Subsequently, regression models are established. Finally, the optimization problem is solved via multi-objective genetic algorithm.

## II. SPRING DESIGN AND STATEMENT OF PROBLEM

Figure 1 demonstrates a schematic model of a so-called compliant tension spring (CTS). The proposed CTS's structure is made by a series of leaf hinges. Entire structure of the proposed spring is monolithically fabricated. It's prototype can be manufactured by wire electrical discharged machining, CNC machining, or 3D printing, and so forth. In order to reduce the stress concentration, corners of the CTS are filleted with radius,  $R$ . The CTS is designed to be subjected a tension load. As seen in Fig. 1.a, by exerting a force,  $F$ , the spring is deformed along the horizontal direction to generate a displacement ( $\Delta x$ ) in the x-axis and another displacement in the y-axis. The fact is that the displacement of the CTS along the y-axis is relatively small, and it is ignored in this study.

In the present article, the CTS is motivated to apply to manipulator, soft robotics and assistive technology. The CTS plays a role as a tension spring, and the main movement of CTS is along the x-axis. Therefore, the displacement,  $\Delta x$ , is considered as a the first fitness function. Besides, the CTS must ensure that the resulting stress should lower than a allowable stress of material to avoid a plastic deformation. So, the stress is considered as the second fitness function.

Dimensions of the CTS are shown in Table 1. Al material is chosen for the CTS due to its lightweight.

TABLE I. GEOMETRICAL PARAMETERS

Parameters	Symbol	Unit: mm
Filleted radius	$R$	0.5
Height of spring	$H$	$26 \leq H \leq 32$
Thickness of flexure hinge	$T$	$2 \leq T \leq 2.5$
Length of flexure hinge	$L_1$	$13 \leq L_1 \leq 17$
Length of vertical link	$L_2$	9
Width of spring	$W$	10
Total length of spring	$L$	90.5



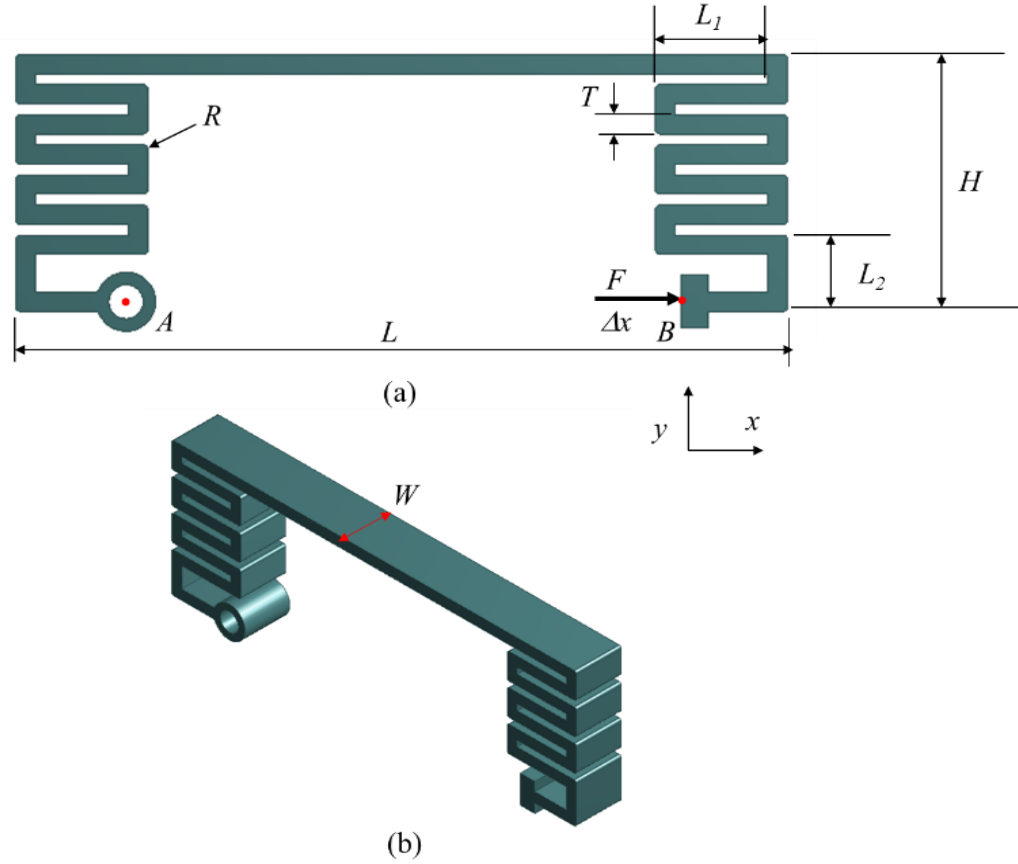


Fig. 1. Compliant tension spring: (a) 2D view and (b) 3D view

The problem of spring design is to generate a large displacement along the tension direction, i.e., a large displacement is desired for a lot of industrial applications. According to the field of compliant mechanism [9, 10], the displacement can be enhanced by implementing geometrical parameters of the CTS. It is well-known that the CST only works in an elastic limit of a material, it is thus easy to appear failures such as plastic deformation, short fatigue life, crack, and fracture. So, the equivalent stress in the CTS should be small to improve overall the performances of spring. Another problem, the displacement is always conflicted with the stress. Therefore, a multiple-objective problem for the CTS is motivated in the present article. The optimization problem for the CTS is stated as follows:

Find the design variables:  $\mathbf{X} = (T, L_1, H)$

The first objective is to maximize the displacement:

$$F_1(\mathbf{X}) = F(T, L_1, H) \quad (1)$$

The second objective is to minimize the equivalent stress:

$$F_2(\mathbf{X}) = F(T, L_1, H) \quad (2)$$

The design constraints are presented as below:

$$\begin{cases} F_1(\mathbf{X}) \geq 20 \text{ mm} \\ F_2(\mathbf{X}) \leq \sigma_y \end{cases} \quad (3)$$

The upper bound and lower bound of the design variables are as:

$$\begin{cases} 2 \text{ mm} \leq T \leq 2.5 \text{ mm} \\ 13 \text{ mm} \leq L_1 \leq 17 \text{ mm} \\ 26 \text{ mm} \leq H \leq 32 \text{ mm} \end{cases} \quad (4)$$

where  $F_1(\mathbf{X})$  and  $F_2(\mathbf{X})$  are the displacement and stress, respectively. The key design parameters include  $T$ ,  $L_1$ , and  $H$ . The remain parameters are constant.  $\sigma_y$  is the yield strength of Al material (280 MPa).

### III. METHODOLOGY

In order to the multiple-objective optimization problem for the CTS, a flowchart of optimization process is presented, as given in Fig. 2. The stepwise procedure is briefly described as follows: (i) An initial design of the CTS begins with a mechanical architecture. (ii) Determination of key design parameters and their upper and lower bounds. (iii) Determination of objective functions and constraints. (iv) Creation of a draft 3D model of the CTS. (v) Finite element analysis is implemented to collect the numerical dataset. (vi) Formulation of regression surface by using Kriging metamodel technique. (vii) Convergent verification of estimated results through regression models. (viii) Verification of the performances indexes. (ix) Implementation of optimization problem by using multi-objective genetic algorithm (MOGA). (x) Evaluation of optimized results.

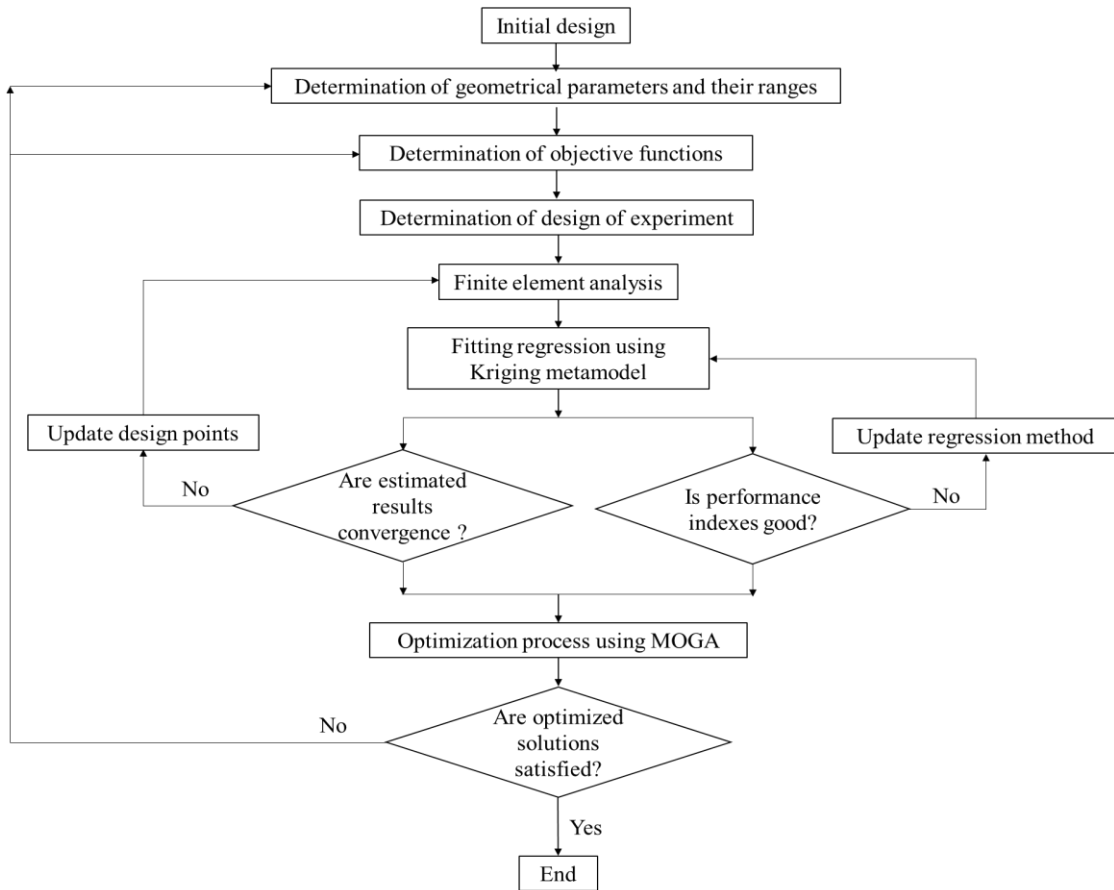


Fig. 2. Flowchart of optimization method.

During the optimization process, it is noted that if the estimated results or the performance indexes are not satisfied, the optimization process is returned to update the design points or change another regression method. Otherwise, the process is continued to move to the MOGA. At this step, if the optimized solutions are not satisfied, the process is returned back the adjustment the range of design variables or consideration the objective functions and constraints. Otherwise, the process is stopped herein.

#### IV. RESULTS AND DISCUSSION

First of all, a draft 3D model of the CTS is created for the finite element analysis. And then, numerical dataset is collected by central composite design and simulations, as given in Table 2. In this step, the CCD technique is utilized to build a plan of experiments.

TABLE 2. NUMERICAL DATASET

No.	$H$ (mm)	$T$ (mm)	$L_I$ (mm)	Displacement (mm)
1	29	2.25	15.025	10.75120831
2	26	2.25	15.025	8.741090775
3	32	2.25	15.025	13.07178307
4	29	2	15.025	15.63985252
5	29	2.5	15.025	7.712244034
6	29	2.25	13	10.13850403
7	29	2.25	17.05	11.41303444
8	26.5609	2.046742	13.37861	11.68915844
9	31.4391	2.046742	13.37861	16.21310234

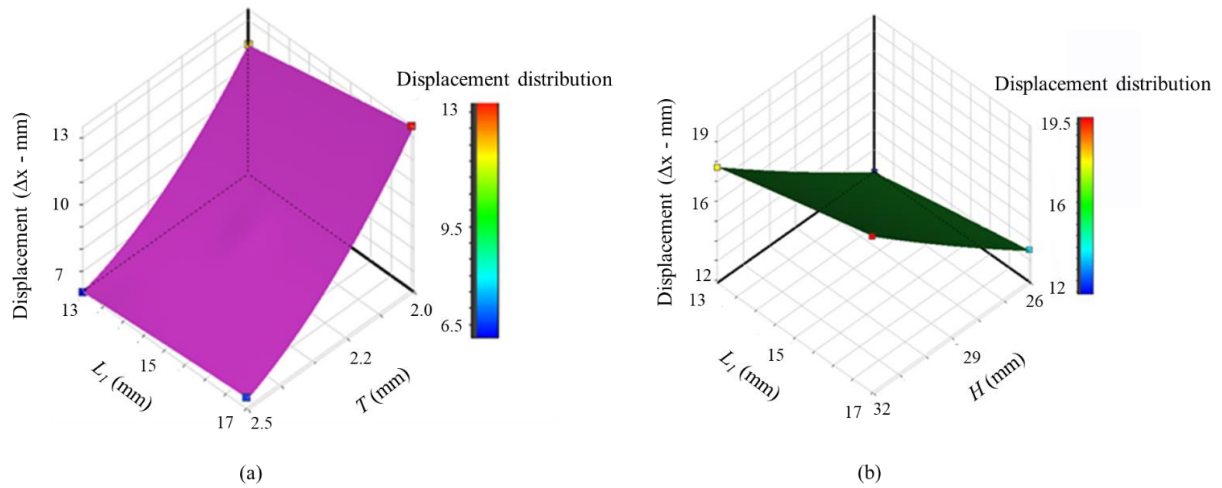
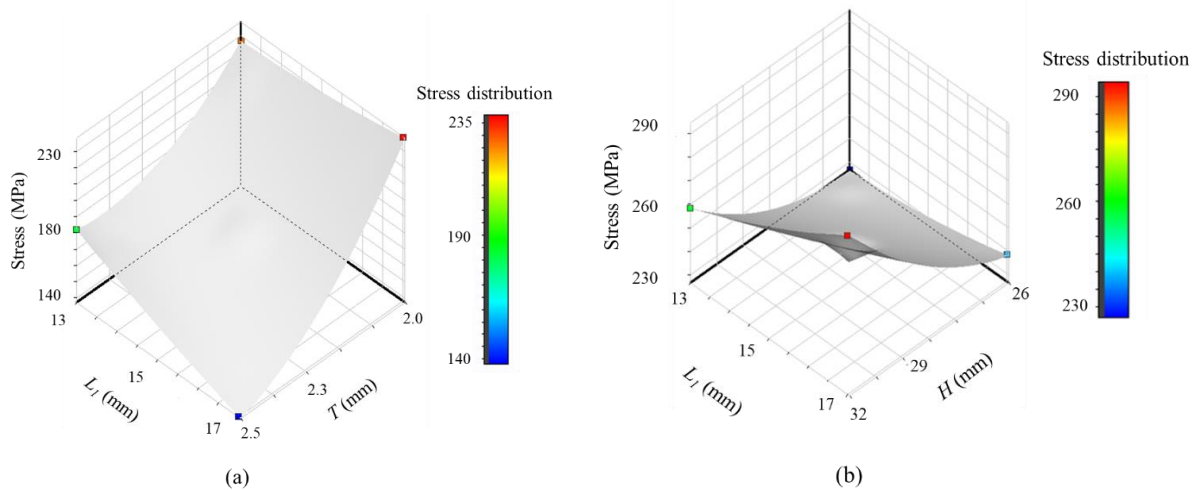
No.	$H$ (mm)	$T$ (mm)	$L_I$ (mm)	Displacement (mm)
10	26.5609	2.453258	13.37861	6.563869476
11	31.4391	2.453258	13.37861	9.123748779
12	26.5609	2.046742	16.67139	12.87727928
13	31.4391	2.046742	16.67139	17.85387802
14	26.5609	2.453258	16.67139	7.160189629
15	31.4391	2.453258	16.67139	9.876755714

Based on the datasets in Table 2, a regression models for the displacement and equivalent stress are formulated through the Kriging metamodel. The results of performance indexes indicated that the developed models are suitable for both performances, as shown in Table 3.

TABLE 3. PERFORMANCE INDEXES.

Index	Symbol	Value
Coefficient of Determination	$R^2$	1
Root Mean Square Error	$RMSE$	1.7376E-09
Relative Maximum Absolute Error	$RMAE$	0
Relative Average Absolute Error	$RAAE$	0

Figure 3 demonstrates the sensitivity of design variables to the displacement. It found that the thickness,  $T$ , and the length,  $L_I$ , strongly contribute to the displacement. Figure 4 shows the sensitivity of design variables to the stress. It is noted that the thickness,  $T$ , and the length,  $L_I$ , also main contribute to the stress.

Fig. 3. Plot of sensitivity analysis of displacement: (a)  $T$  and  $L_I$ , (b)  $L_I$  and  $H$ Fig. 4. Plot of sensitivity analysis of stress: (a)  $T$  and  $L_I$ , (b)  $L_I$  and  $H$ 

At last, by the use of MOGA algorithm, the optimal results showed that the displacement and the equivalent stress are found about 32.51 mm and 258.35 MPa. The optimal parameters are determined with  $H$  of 32 mm,  $T$  of 2 mm, and  $L_I$  of 13 mm. The displacement and stress are satisfied with the initial constraints, as given in Table 4.

TABLE 4. OPTIMAL RESULTS

Optimal solution	Value
$H$ (mm)	32
$T$ (mm)	2
$L_I$ (mm)	13
Displacement (mm)	32.51
Stress (MPa)	258.35

## V. CONCLUSIONS AND FUTURE REMARKS

In this paper, a computational optimization process for the CTS spring is proposed. The CTS's structure consists of

rectangular hinges which are arranged in a series. The displacement and the equivalent stress are considered as two objective functions. Meanwhile, the geometrical parameters are taken as design variables.

A draft 3D model of the CTS is designed, the CCD technique is applied to establish the experimental plan. And then, FEA simulations are carried out. Based on the numerical data, the regression models for both objective functions are formulated by using the Kriging metamodel. Finally, the optimal solutions are found through MOGA.

The optimized results found that the displacement and stress are approximately 32.51 mm and 258.35 MPa. These values are satisfied with the design goals and constraints. MOGA is an efficient tool to reach the Pareto-optimal set for the CTS.

Future research will focus on manufacturing a prototype, and a verification of predicted values is performed by physical experiments.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 107.01-2019.14.

REFERENCES

- [1] Howell LL, Magleby SP, Olsen BM.: Handbook of Compliant Mechanisms, John Wiley & Sons, (2013).
- [2] Wang R, Zhang X.: Optimal design of a planar parallel 3-DOF nanopositioner with multi-objective. *Mech Mach Theory* 112:61–83 (2017).
- [3] Lobontiu N.: Modeling and design of planar parallel-connection flexible hinges for in- and out-of-plane mechanism applications. *Precis Eng* 42:113–132 (2015).
- [4] Sakhaei AH, Kaijima S, Lee TL, et al.: Design and investigation of a multi-material compliant ratchet-like mechanism. *Mech Mach Theory* 121:184–197 (2018).
- [5] Chau N Le, Dao TP, Nguyen VTT.: Optimal Design of a Dragonfly-Inspired Compliant Joint for Camera Positioning System of Nanoindentation Tester Based on a Hybrid Integration of Jaya-ANFIS. *Math Probl Eng* 2018 (2018).
- [6] Dong W, Chen F, Gao F, et al.: Development and analysis of a bridge-lever-type displacement amplifier based on hybrid flexure hinges. *Precis Eng* 54:171–181 (2018).
- [7] Thoai N, Ngoc T, Chau L.: A New Butterfly-Inspired Compliant Joint with 3-DOF In-plane Motion. *Arab J Sci Eng*. 2020 (2020).
- [8] Chau N Le, Tran NT, Dao T.: Design and Performance Analysis of a TLET-Type Flexure Hinge. 2020 (2020).
- [9] Wu J, Zhang Y, Cai S, Cui J.: Modeling and analysis of conical-shaped notch flexure hinges based on NURBS. *Mech Mach Theory* 560–568 (2018).
- [10] Hao G, He X, Awtar S.: Design and analytical model of a compact flexure mechanism for translational motion. *Mech Mach Theory* 142:103593 (2019).

# Application of Fuzzy Control Algorithm to Start a Large -Capacity Synchronous Motor

Quoc Hung Duong  
Department of Electrical Engineering  
Thai Nguyen University of Technology  
Thai Nguyen city, Vietnam  
quochungkd@tnut.edu.vn

Huu Cong Nguyen  
Board of Directors  
Thai Nguyen University  
Thai Nguyen city, Vietnam  
conghn@tnu.edu.vn

The Cuong Nguyen  
Board of Directors  
ASO Mechatronics Joint Stock Company  
Song Cong City, Vietnam  
thecuong@aso.com.vn

Hong Quang Nguyen  
Department of Automation  
Thai Nguyen University of Technology  
Thai Nguyen city, Vietnam  
quang.nguyenhong@tnut.edu.vn

**Abstract**— Large-capacity synchronous motors are widely used in industry because of their advantages. However, it has a complicated structure, and especially it has the excitation controller at the rotor side. In the start-up mode, it is crucial to choose the right time to supply DC excitation source to the rotor winding. This helps the motor may start more smoothly, reduce the stator current when starting, and improve the life of mechanical structures. This article presents the application of the fuzzy control algorithm for the excitation controller to improve the starting quality of the large capacity synchronous motor. The simulation results and experiment results on the sizeable synchronous motor proved the correctness of the algorithm.

**Keywords**—Excitation; Fuzzy control; Model of synchronous motor; Matlab Simulink

## I. INTRODUCTION

Synchronous motors and induction motors are the most widely used types of AC motor. The difference between the two types is that the synchronous motor rotates synchronously with the grid frequency, In contrast, induction motors have a sliding coefficient, The rotor speed is a bit slower than the stator magnetic field speed, to develop torque [1-6]. Synchronous motors are available from small capacity types, self-excitation, to broad capacity types with external excitation source. Small synchronous motors are used in timing applications such as in synchronous clocks, timers in appliances [7-8], tape recorders, and precision servomechanisms [9-11] in which the motor must operate at a precise speed. The large synchronous motor used in industry, it provides two essential functions. First, it is a highly efficient means of converting AC energy to work [12-15]. Second, it can operate as a power factor regulator to improve the power factor of the grid [16-19].

Starting a large capacity synchronous motor is more complicated than the asynchronous motor because it is necessary to determine the exact time to apply the DC excitation source to the rotor winding called "catching" synchronous. In the working mode, this DC source must be adjusted to keep the stability of the power factor to improve motor performance. Therefore synchronous motors require higher operating costs than asynchronous motors [20-22].

Starting the motor by measuring the rotor speed method was also mentioned in these works [1], [2], [4], [28]. At the beginning of the start-up mode, the excitation source is

disconnected. The motor starts as an induction motor. At low speeds with high slip coefficients, a high voltage generated in the field winding can damage the motor. Therefore the field winding is shorted through the field resistance to disperse this voltage. Besides this field resistance adds to motor starting torque. As the motor reaches close to synchronous speed, the field is energized from dc supply. This method is employed for no load or low load starting.

The auxiliary motor is used to bring the synchronous motor near synchronous speed [3]. When motor speed is closed to the synchronous speed, the field winding is connected to a DC excitation source. The field of stator locks the field of the rotor, and the motor starts running at synchronous speed. At the end of the start-up mode, the auxiliary motor can be disconnected from the supply.

Similar to speed measurement, rotor frequency measurement is also applied [29], [30], the timing of the synchronized catch is when the Rotor frequency is about 2 - 4Hz. Some other researchers find the timing of the synchronized catch when the stator current is minimum [31], [32]. The excitation current builds up relatively slowly when excitation voltage is applied.

Most of the research mainly focuses on Rotor speed or Rotor frequency without paying attention to the phase angle of the rotor current, so sometimes the application of DC excitation is reversed-phase, and the phenomenon of current conflict occurs on the stator side, that causes vibration when startup with a massive load. This article presents an optimal algorithm for the excitation controller of a large synchronous motor. The fuzzy logic algorithm is used for "catching" synchronous in the start-up mode. This helps the motor may start more smoothly, reduce the stator current when starting, and improve the life of mechanical structures.

## II. RESEARCH METHOD

For this paper, the fuzzy logic algorithm is used for "catching" synchronous in the start-up mode. An experimental model was built for an asynchronous motor with a capacity of 500kw applied at a pump station.

### A. Fuzzy Controller Design

Today, the most large synchronous motor is the salient pole type with utilizing squirrel-cage windings in the pole faces of the synchronous motor rotor [19- 27]. These



windings allow for a reaction (or acceleration) torque to be developed in the rotor as the AC supplied stator windings induce current into the squirrel cage windings (Fig. 1). Therefore, in the beginning, the synchronous motor starts as an induction motor.

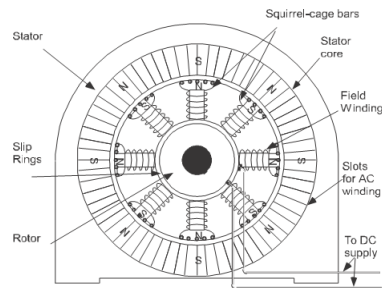


Fig. 1. Salient pole Synchronous motor

At the beginning of the start-up, the DC excitation source is disconnected, and the discharger resistors are applied to the Rotor circuit to disperse the large electromotive induction that can damage the Rotor windings. The motor is started as an induction motor with a sliding coefficient decreasing from 1. When the motor starts to approach synchronous speed, the slip coefficient decreases toward 0. If  $I_s$  is stator current and  $I_{FD}$  is rotor current, the induction current from the Stator side, then  $I_s$  would be about  $180^\circ$  from rotor current  $I_{FD}$ , and the flux would be  $90^\circ$  behind  $I_{FD}$ . The point of maximum-induced flux ( $\emptyset$ ) occurs as the rotor current  $I_{FD}$  passes through zero from negative to positive. See Fig. 2.

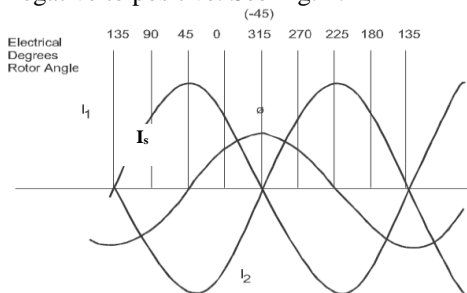


Fig. 2. Typical transformer rotor flux and current (constant slip)

The angle of the rotor at which  $I_s$  and  $I_{FD}$  go through zero depends upon the reactance to resistance ratio in the field circuit. There is a value of the reactance to resistance ratio when the angle of the rotor shifts toward  $-90^\circ$ . At high speed or low slip and low frequency, reactance decreases, and the angle shifts toward  $0^\circ$  if the high value of resistance is put into the field circuit. As the stator goes beyond  $-45^\circ$ , the torque increases. At this point, The rotor current yields a convenient indicator of maximum flux and increasing torque from which excitation is applied for maximum effectiveness. The fuzzy control algorithm is applied to external excitation in the correct polarity to inc  $I_{FD}$  this trapped flux at this instant makes maximum use of its existence. At this point, the stator pole has just moved by and is in position to pull the rotor forward into synchronous alignment. See Fig.3 and Fig.4.

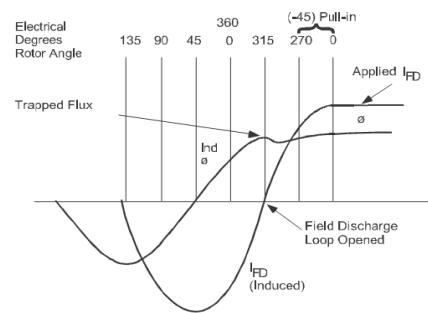


Fig. 3. Typical rotor flux and current at pull-in

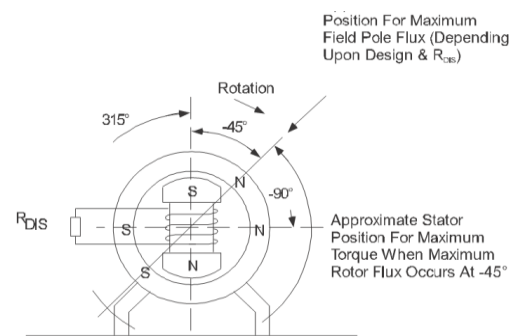


Fig. 4. Angular displacement of Rotor

The system control structure diagram is shown in Fig.5. This figure shows that the necessary and sufficient conditions to "catch" synchronously are: Rotor speed reaches nearly synchronizing speed (about 90% to 95% of synchronous speed), and the rotor current flow through 0 in a positive direction, corresponding to Rotor angle is  $-45^{\circ}$ . The fuzzy control algorithm is applied to the controller to optimize the starting process and is built according to the MISO structure with 3 input signals (Fig. 6): Rotor speed  $[n]$ , Current Rotor  $[I(pu)]$  and its derivative  $[DI(pu)]$ . The output signal of the controller determines the time at which the DC excitation source is supplied to rotor windings.

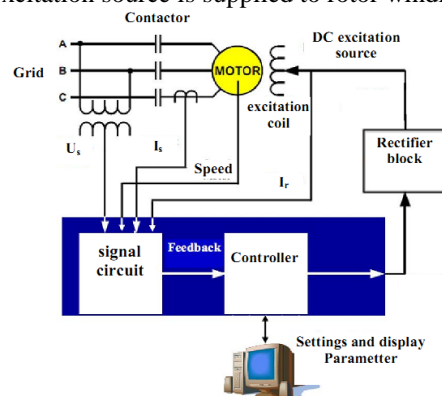


Fig. 5. System control structure

The membership functions are trimf shape. The rules of the controller are implemented according to Max - Min, Defuzzification is done by the height method. The inputs of the fuzzy controller are shown as fig. 7 to fig. 9 and the output of the fuzzy controller is shown as fig. 10. The Diagram of simulation of System in Matlab Simulink is shown as fig. 11.

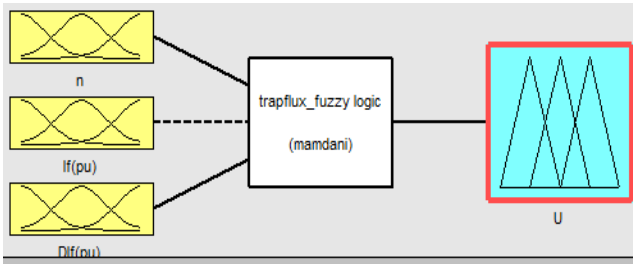


Fig. 6. Structure diagram of the fuzzy controller

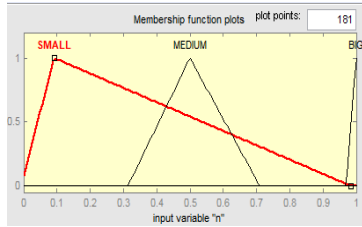


Fig. 7. Rotor speed membership functions

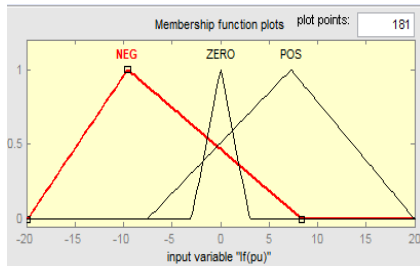


Fig. 8. Rotor current membership functions

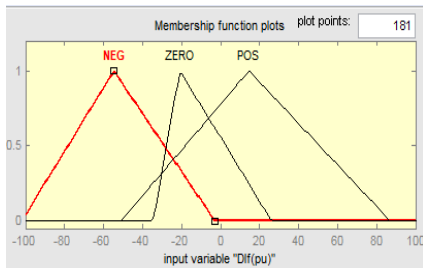


Fig. 9. Rotor current derivative membership functions

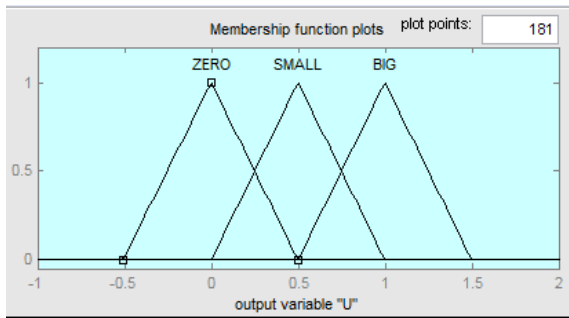


Fig. 10. Signal output membership functions

The mamdani rulers of the fuzzy controller:

If  $n = \text{"SMALL"}$  and  $If(pu) = \text{"xxx"}$  and  $Dif(pu) = \text{"xxx"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"MEDIUM"}$  and  $If(pu) = \text{"NEG"}$  and  $Dif(pu) = \text{"xxx"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"MEDIUM"}$  and  $If(pu) = \text{"ZERO"}$  and  $Dif(pu) = \text{"xxx"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"MEDIUM"}$  and  $If(pu) = \text{"POS"}$  and  $Dif(pu) = \text{"NEG"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"MEDIUM"}$  and  $If(pu) = \text{"POS"}$  and  $Dif(pu) = \text{"ZERO"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"MEDIUM"}$  and  $If(pu) = \text{"POS"}$  and  $Dif(pu) = \text{"POS"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"BIG"}$  and  $If(pu) = \text{"NEG"}$  and  $Dif(pu) = \text{"xxx"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"BIG"}$  and  $If(pu) = \text{"ZERO"}$  and  $Dif(pu) = \text{"NEG"}$  then  $U = \text{"ZERO"}$

If  $n = \text{"BIG"}$  and  $If(pu) = \text{"ZERO"}$  and  $Dif(pu) = \text{"POS"}$  then  $U = \text{"BIG"}$

If  $n = \text{"BIG"}$  and  $If(pu) = \text{"POS"}$  and  $Dif(pu) = \text{"NEG"}$  then  $U = \text{"SMALL"}$

If  $n = \text{"BIG"}$  and  $If(pu) = \text{"POS"}$  and  $Dif(pu) = \text{"ZERO"}$  then  $U = \text{"SMALL"}$

If  $n = \text{"BIG"}$  and  $If(pu) = \text{"POS"}$  and  $Dif(pu) = \text{"POS"}$  then  $U = \text{"BIG"}$

In which:

+ "xxx" can be "NEG" or "ZERO" or "POS"

+ If  $n$  is "SMALL" or "MEDIUM", then no matter what  $If(pu)$  and  $Dif(pu)$  are, there is the same  $U$  and it is "ZERO".

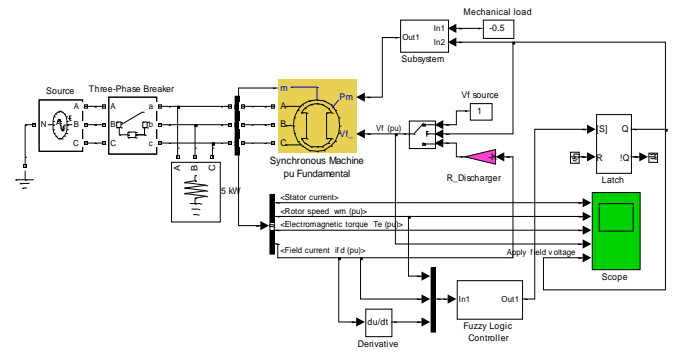


Figure 11. Diagram of simulation on Matlab Simulink

### B. Experimental model Design

Experimental model designed to control the synchronous pump motor with a capacity of 500kw. The diagram of the experimental system is shown in Fig. 12.

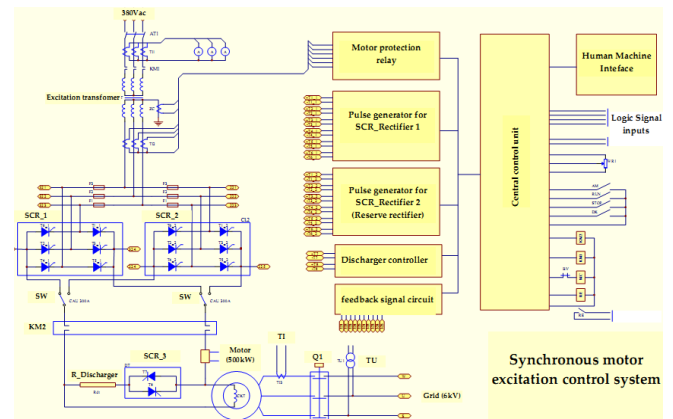


Fig. 12. General diagram of the experimental system model

Where:

+ **Synchronous motor:** The motor is used for pumping stations with parameters shown as table I and fig.13

TABLE I. THE MOTOR'S PARAMETERS OF EXPERIMENTAL SETUP

Parameter	Value	Parameter	Value
Type	salient pole	Rated current	58.3A
Model	TNPBC 16-41-20TN	Speed	300 R/m
Capacity	500kW (605KVA)	Rated current excitation	227
Rated voltage	6kV	Rated voltage excitation	57V



Fig.13. The synchronous motor (500kW)

+ **SCR\_1 and SCR\_2:** The authors use a 3-phase bridge rectifier to generate the DC excitation source. In which 1 set of activities and a set to reserve. The switch (SW) is used to select the working rectifier.

+ **R\_Discharger (the field resistance) and SCR\_5:** When start-up, a high voltage would be induced in the field winding. It usually has a large number of turns, which can damage it. In fact, to avoid high starting current in the field resistance, it is shorted through resistance several times the field resistance adds to motor starting torque. The SCR\_3 is controlled by a discharger controller.

+ **Pulse generator for SCR\_Rectifier 1 and 2** are used to generate pulses for SCR\_1 and SCR\_2.

+ **Feedback signal circuit:** These signals are sent to the central controller to select the time of the synchronized catch. and control the DC excitation source during operation.

+ **The central controller** is programmed according to algorithms to start the motor and optimize the working process.

+ **Human Machine Interface** is used to set and display working parameters

Experimental images are shown in Fig.14 to Fig.16.



Fig.14. The experimental system model

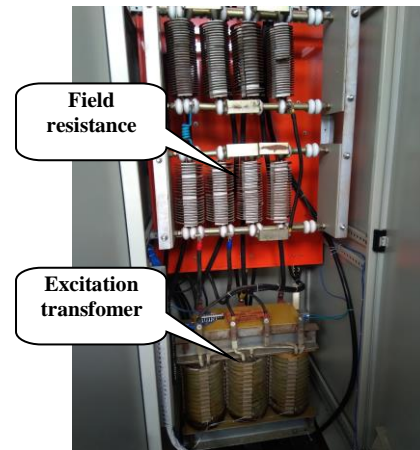


Fig.15. The field resistance and excitation transformer

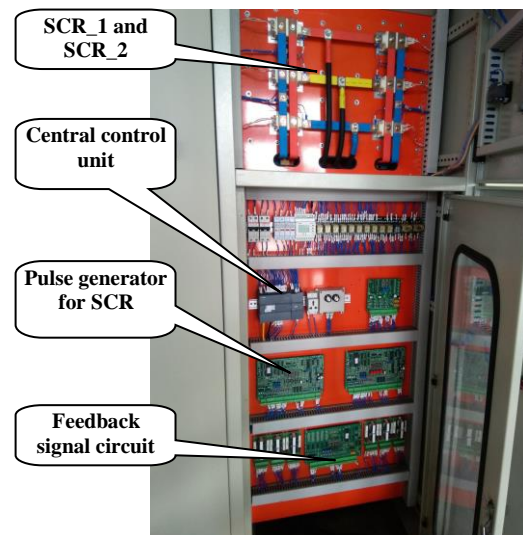


Fig.16. The control cabinet

At the beginning of starting, the contactor KM2 are closed, the rectifier SCR\_1 (or SCR\_2) has not been opened. The circuit breaker Q1 closes to supply power to the stator. At the same time, SCR\_3 opens to apply the field resistance. When the motor speed reaches 95% of the synchronous speed, or the frequency of the rotor current is approximately 4Hz, it is a necessary condition to "catch"

synchronously. The feedback signal circuit determines the phase angle of the Rotor current at the time of the maximum induced flux value. At this time, the excitation system will be applied and the field resistance will be disconnected, the motor will start up to the rated speed. The excitation system is applied on the exact synchronization between the rotation frequency of the motor and the frequency of the grid to ensure that there is no sudden surge of current in the rectifier bridge causing damage to the Thyristor and the system.

### III. RESULTS AND ANALYSIS

#### A. Simulation results

The synchronous motor parameters used in this simulation are listed in table II.

TABLE II. THE SYNCHRONOUS MOTOR'S PARAMETERS USED IN THE SIMULATION

Parameter	Notation	Value	Unit
Stator	$R_s$	0.03788	pu
	$L_l$	0.08	pu
	$L_{md}$	2.16	pu
	$L_{mq}$	0.94	pu
Dampers	$R_{kd}$	0.1463	pu
	$L_{lkd}$	0.30485	pu
	$R_{kq1}$	0.05754	pu
	$L_{lkq1}$	0.05281	pu
Field	$R_f$	0.02124	pu
	$L_{lfd}$	0.1719	pu
The initial value of the current and phase	$i_a$	0	pu
	$i_b$	0	pu
	$i_c$	0	pu
	$ph_a$	0	deg
	$ph_b$	0	deg
	$ph_c$	0	deg
DC Excitation source	$V_f$	1	pu

Simulation results are in Fig. 17 to Fig. 21.

Fig. 17 and fig. 18 show that when the fuzzy algorithm is not used, it only bases on rotor speed. The apply of the excitation source to the rotor at the wrong time will cause the stator current to oscillate, the electromagnetic torque to be small and also to oscillate.

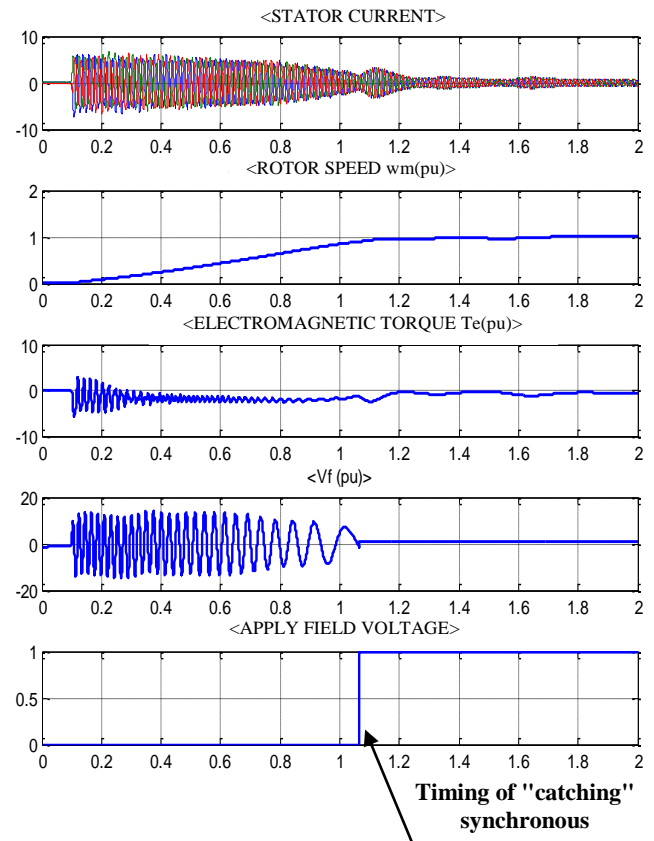


Fig.17. Simulation results with no fuzzy algorithm at 50% load and rotor speed reach 85% of synchronous speed

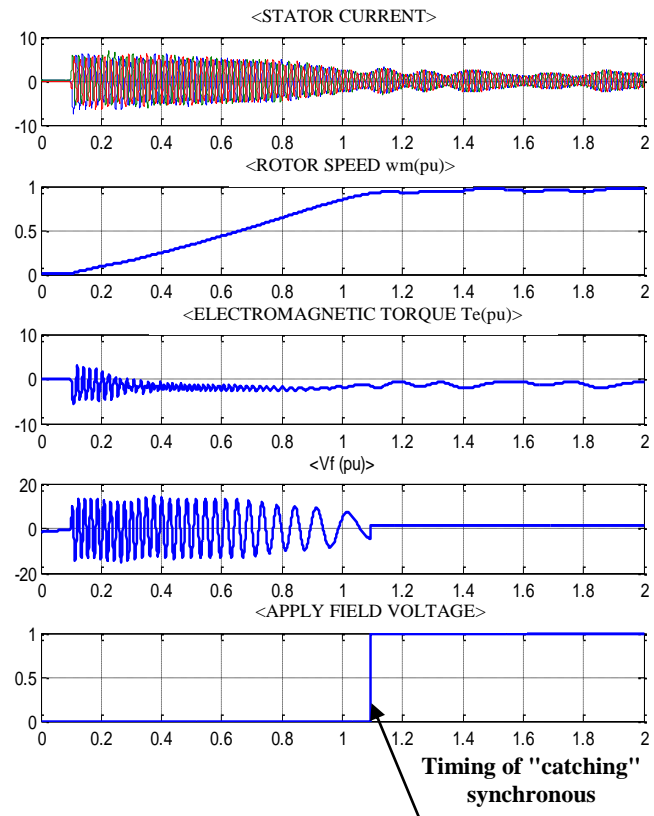


Fig.18. Simulation results with no fuzzy algorithm at 100% load and rotor speed reach 93% of synchronous speed



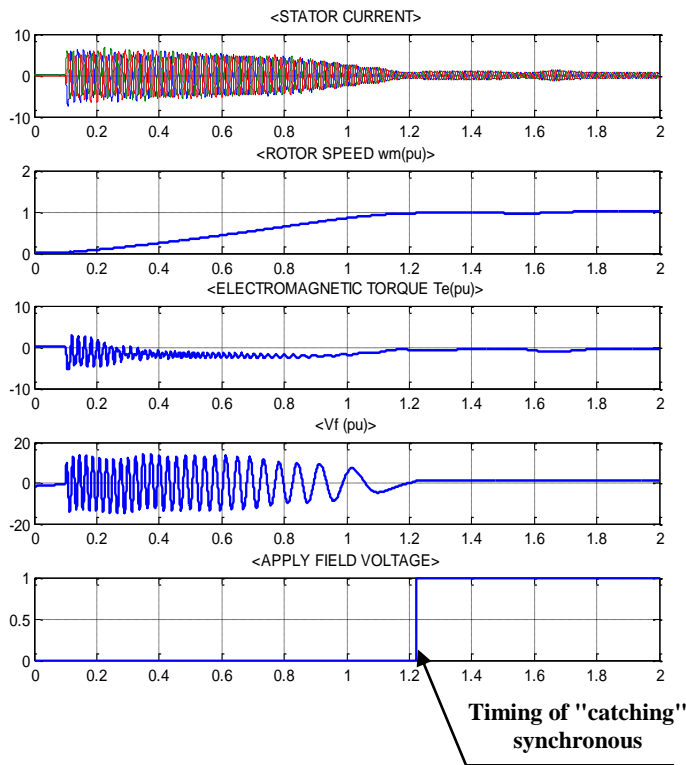


Fig.19. Simulation results with fuzzy algorithm at 50% load

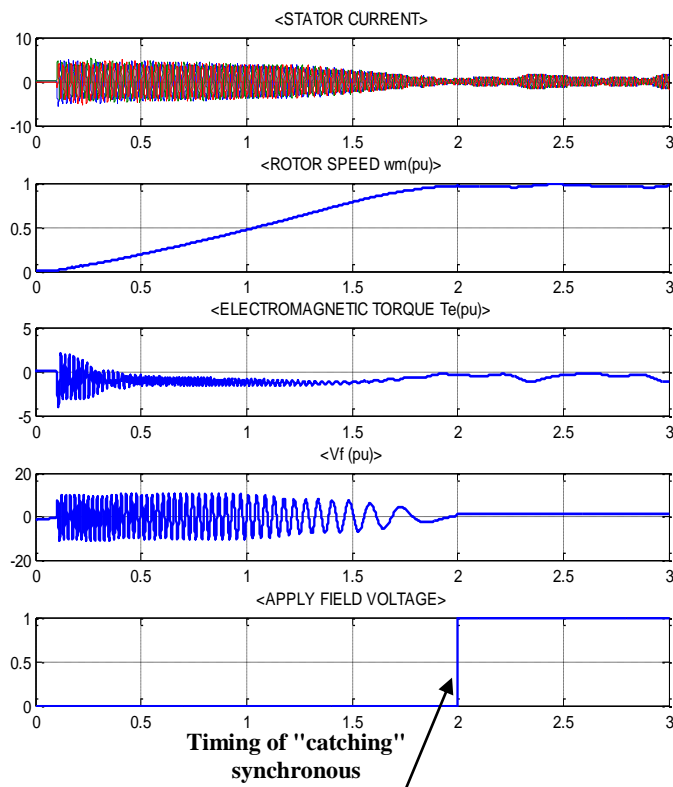


Fig.20. Simulation results with fuzzy algorithm at 100% load

Fig .19 and fig.20 show that the fuzzy control algorithm used has determined the exact time of the synchronized catch at the speed of  $\geq 95\%$  of the rated speed and the Rotor current through point 0 in the positive direction. At this time, the induced flux of the rotor reaches the maximum, thus the torque has the greatest value.

When the motor carries the load, the time of synchronized catch is determined by about 1.22 seconds with 50% of the load (Fig.19), and about 2 seconds with 100% of the load (Fig.20). In this case, there is no conflict in the current, Stator current is stable, the speed is stable and the motor is started smoothly.

### B. Experimental results

The algorithm is built and programmed on fast processing chips. Pre-closed contactor KM2 to reduce delay due to mechanical switching. Controlling pulses for Thyristors are sent by pulse generators to open the rectifier bridges to increase the response speed when synchronizing.

In addition to using the Encoder to measure the speed of the motor, the authors also use electrical circuits measuring the frequency of the rotor current to ensure redundancy. Experimental results show that the timing of the synchronized catch is about 3 to 4s; the motor starts smoothly; there is no vibration and is stable at rated speed (fig. 21). The excitation voltage is about 39Vdc (fig. 22); the excitation current is about 171A (fig. 22). The frequency of rotor current when synchronizing is about 4hz (fig. 25).

The test results showed that:

- There is no failure during the start-up mode;
- The motor starts very smoothly, the time of start-up from 3 to 4s (depending on the load);
- There is no surge in the Stator current at the timing of "catching" synchronous.

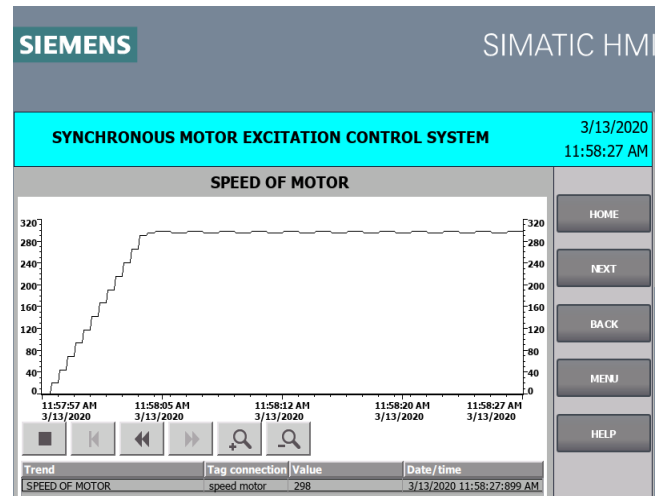


Fig. 21. Speed of Motor

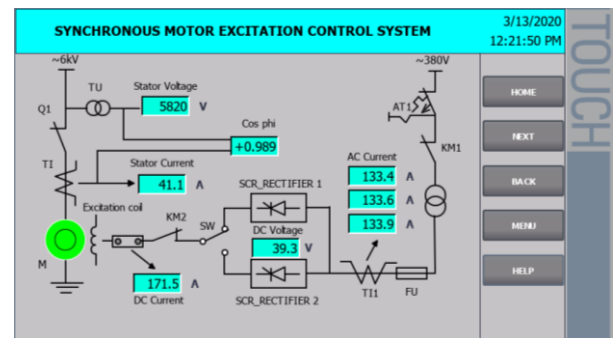


Fig. 22. Structure of the system in HMI



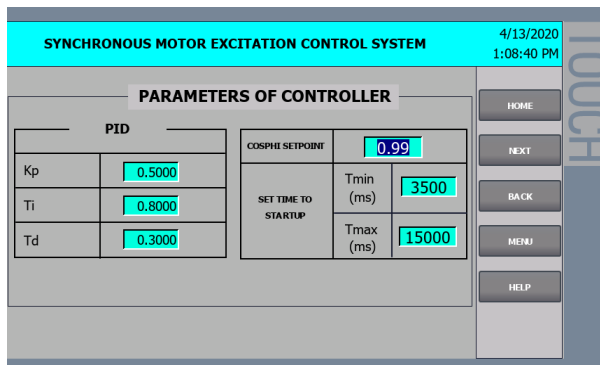


Fig. 23. Parameters of controller

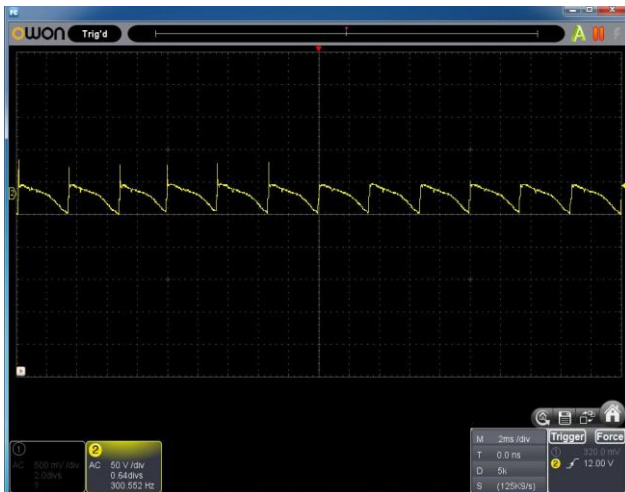


Fig. 24. Voltage of rectifier SCR\_1

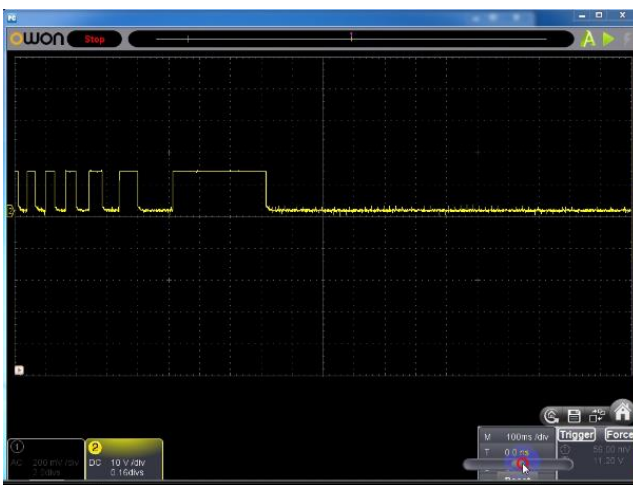


Fig. 25. Frequency of rotor current when synchronizing

#### IV. CONCLUSIONS

Synchronous motors have many advantages compared to asynchronous motors with the same power, but it has a complex structure. Hence it is more difficult to control it in both the starting and working process. The paper added an algorithm to catch sync. The synchronous "catching" time is determined when the rotor speed is close to the synchronous speed and the induced flux of the rotor reaches the maximum, thus the torque has the greatest value. This helps the motor may start more smoothly, reduce the stator current when starting. It can help to increase the service life of

engine and mechanical structure. Simulation and experimental results proved the correctness of the algorithm.

#### DATA AVAILABILITY

The practical data of the road profile used to support the findings of this study have been collected, measured, and processed by the research group under grant no. B2018-TNA-58. The road profile data used to support the findings of this study are available from the corresponding author upon request.

#### ACKNOWLEDGMENT

This research was funded by the Vietnam Ministry of Education and Training under grant no. B2018-TNA-58, and ASO Mechatronics Joint Stock Company supported resources and equipment for this project.

#### REFERENCES

- [1] D P Kothari and I J Nagrath, "Electric machine", TMH Publishers second edition, Pp 250-260, 1998.
- [2] M. G. Say, "Performance and design of ac machine", CBS Publishers third edition, Pp250-260, 2002.
- [3] DR. P.S. Bimbrha, "Electrical Machinery", Khanna Publisher seventh edition, Pp676-680, 2009.
- [4] Irving L. Kosow, "Electric machinery and Transformer", PHI Learning private ltd. second edition, Pp218-225, 2008.
- [5] P.K. Mukherjee and S. Chakraverti, "Electrical machine", Dhanpat Rai Publishers fifteenth edition, Pp635-640, 2004.
- [6] H. Macbahi, A. Ba-razzouk, J. Xu, A. Cheriti, and V. Rajagopalan; "A unified method for modeling and simulation of three phase induction motor drives", 2000.
- [7] S. Yamamura; "Saliency torque and V-curve of permanent-magnet-excited synchronous motor", International conference on Unconventional Electromechanical and Electrical Systems, Russia, June, 2000.
- [8] Enrique L. Carrillo Arroyo, "Modelling and simulation of permanent magnet synchronous motor drive system", M. Sc. Thesis, University of Puerto Rico, 2006.
- [9] S. Onoda and A. Emadi, "PSIM-based modelling of automotive power systems: conventional, electric, and hybrid electric vehicles", Vehicular Technology, IEEE Transactions on, 2004, vol. 53, pp. 390-400.
- [10] Mostafa.A. Fellani, and Dawo.E. Abaid, "Sliding Mode Control of Synchronous Reluctance Motor", International Journal of Electronics, Circuits and Systems, 2004, Vol.3, No.2, 2009.
- [11] David ocn, "Direct torque control of a permanent magnet synchronous Motor", Master's Degree Project Stockholm, Sweden 2005.
- [12] J.C. Pequeña, E. Ruppert and M.T. Mendoza; "On the Synchronous Generator Parameters Determination Using Dynamic Simulations Based on IEEE Standard", Industrial Technology (ICIT), IEEE International Conference on, Viña del Mar, Chile, 2010, pp. 386-391.
- [13] Damir Suminal, Gorislav Erceg, Tomislav Idzotic, "Comparison of the excitation control of a synchronous generator with fuzzy logic controller and PI voltage controller"; Faculty of Electrical Engineering and Computing Unska 3, Zagreb, Croatia, 2012.
- [14] IEEE Guide, "Synchronous Generator Modeling Practices and Applications in Power System Stability Analyses", IEEE Std 1110-2002 (Revision of IEEE Std 1110-1991 [2003]).
- [15] IEEE 1110, "Guide for Synchronous Generator Modeling Practices and Applications in Power System Stability Analyses", 2002.
- [16] YuQi Zhang, "Advanced synchronous machine modeling", Teses and Dissertations-Electrical and Computer Engineering. 118. [https://uknowledge.uky.edu/ece\\_etds/118](https://uknowledge.uky.edu/ece_etds/118), 2018.
- [17] Moeini, A., et al, "Synchronous Machine Stability model", an Update to IEEE Std 1110-2002 Data Translation Technique, IEEE standards panel sessions, 2018
- [18] L.Shi-Dong, L.Jian-Zhao; "Analytic calculation of V-curve for salient-pole synchronous electric machine"; Proceedings of the

- Chinese Society of Electrical Engineering, Vol. 28, no. 18, pp. 110-113, 15 June 2008.
- [19] Quoc Hung Duong, "Modeling and simulation of synchronous motors on Matlab – Simulink"; The National Conference on Electronics, Communications and Information Technology REV- Viet Nam, 2016.
- [20] IEEE Transactions on Energy Conversion, "Large Synchronous Machines", Vol. 10, No. 3, September 1995.
- [21] Jim parrish, steve moll, & richard c. schaefer; "Plant efficiency benefits resulting from the use of synchronous motors"; IEEE industry applications magazine, MAR/APR 2006, www.ieee.org/ias, 2006.
- [22] Krause, P.C, "Analysis of Electric Machinery", Section 12.5. New York: McGraw-Hill, 1986.
- [23] IEEE 115, IEEE Guide, "Test Procedures for Synchronous Machines Part I" - Acceptance and Performance Testing Part II-Test Procedures and Parameter Determination for Dynamic Analysis, 1995.
- [24] IEEE Trans, "Power Apparatus and Systems", vol. 96, pp. 1211-1218, July/Aug 1977.
- [25] IEEE 115, "Guide for Test Procedures for Synchronous Machines Part I" - Acceptance and Performance Testing Part II-Test Procedures and Parameter Determination for Dynamic Analysis, 2009.
- [26] Matwork, "Model the dynamics of three-phase round-rotor or salient-pole synchronous machine", 2010.
- [27] Kilowatt classroom, LLC, "Synchronous Motor", 2004.
- [28] Quoc Hung Duong, Huu Cong Nguyen, The Cuong Nguyen; "Starting the large synchronous motor by speed method"; Thai Nguyen University Journal of Science and Technology, T. 225, S. 06 5/2020.
- [29] WEG group, "The ABC's of Synchronous Motors", At www.electricmachinery.com.
- [30] 24. E.C. Bortoni, J.A. Jardini; "A Standstill Frequency Response Method for Large Salient Pole Synchronous Machines", IEEE Trans on E.C, Vol. 19, No. 4, pp. 687-691, December 2004.
- [31] Arun Kumar Datta, "Manisha Dubey, Shailendra Jain; "Modelling and Simulation of Static Excitation System in Synchronous Machine Operation and Investigation of Shaft Voltage"; Hindawi Publishing Corporation Advances in Electrical Engineering, Volume 2014, Article ID 727295, 9 pages, at <http://dx.doi.org/10.1155/2014/727295>
- [32] Bill Horvath, "Synchronous Motors & Sync Excitation Systems", Western Mining Electrical Association, TM GE Automation Systems, 2009.

# Mobile learning in non-English Speaking Countries: Designing a Smartphone Application of English Mathematical Terminology for Students of Mathematics Teacher Education

Bui Anh Tuan

*Department of Mathematics Education,  
Teachers College, Can Tho University  
Can Tho city, Vietnam  
batuan@ctu.edu.vn*

Lam Minh Huy\*

*Department of Mathematics Education,  
Teachers College, Can Tho University  
Can Tho city, Vietnam  
lmhuy1997@gmail.com*

Nguyen Hieu Thanh

*Department of Mathematics Education,  
Teachers College, Can Tho University  
Can Tho city, Vietnam  
thanhb1700039@student.ctu.edu.vn*

Tieu Ngoc Tuoi

*Department of Mathematics Education,  
Teachers College, Can Tho University  
Can Tho city, Vietnam  
tuoi1808299@student.ctu.edu.vn*

Huynh Tuyet Ngan

*Department of Mathematics Education,  
Teachers College, Can Tho University  
Can Tho city, Vietnam  
nganb1900366@student.ctu.edu.vn*

**Abstract**—With the expeditious development of the 4.0 technology era, learning English in non-English speaking countries is more crucial and meaningful than ever. Thus the application of information technology for teaching has been increasingly focused and interested, particularly the application of m-learning to support teaching and learning is also progressively popular. Following this trend, this paper conducts to build the application of learning Mathematics by using the English language for the primary object of Students of Mathematics Teacher Education through three steps: First of all, survey the need for studying Mathematics by using English language and analyze the factors of the affecting to students' English proficiency. After that, plan and design an application for teaching Mathematics by using the English language. Finally, experiment and get opinions. The results obtained in the study show that learners who receive this application are quite positive.

**Keywords**—*Mobile learning, Smartphone application, Mathematics Education, English Education*

## I. INTRODUCTION

In recent years, with the 4.0 industrial revolution and the expeditious development of the digital age, e-learning methods have been promptly developed and become an inevitable trend in education. Especially, mobile learning (m-learning) gradually asserts its position compared to the remaining forms. According to Georgiev et al. (2004) and Wang (2017), m-learning has proved to be superior to other tools because of such advantages: learning is not limited in time and space. In addition, learning tools are mobile devices (computers, smartphones, tablets, etc.), used very ubiquitous in everyday life. It is also easy to interact and exchange with other learners in different places.

The advantage of m-learning is due to the technology boom in general, and smartphones in particular. A research result of Teodorescu (2015) shows that smartphones are used most often with 46% compared to 39% laptops, 9% tablets and 6% with other devices. Also in this study, the author provided an encouraging number (93%) of the percentage of students

who are interested in and consider m-learning to be an effective learning tool. Furthermore, Klimova (2019) pointed out that the majority of students using the mobile application achieved significantly better results than other students on the final exam.

On the other hand, the teaching of Mathematics by using the English language in Vietnam is increasingly focusing and expanding, it is also potential to develop oneself. Nevertheless, at the present, the English ability of students is not high and specific: according to Ministry of Education and Training's statistics (graduating from National High School 2019), it shows that foreign language is one of the two subjects with the lowest average test scores. The average score in English is 4.36 points, and up to 68% of the students are below average. This reflects the current situation of students' ability to learn English, affecting the quality of university entrance students.

Combined with the trend of Mobile Learning, an application is needed to support English lookup and learning. From mobile app stores like CH Play and AppStore, it can be easily found that several related applications, especially some of the most popular and widespread applications such as: Learn English by Listening, Learn English Grammar, English Tenses, English Reading, Practice English Grammar, English Grammar Ultimate, Speak English, English Test Package, and so on. (Klimova, 2019). Nevertheless, there are a few application for students majoring in Mathematics and there is no application supports Vietnamese.

Motivated by this, this paper conducts a study and builds a mobile application to support the study of terminology in Mathematics by using the English language for students majoring in Mathematics. The outline of the paper is presented as follows: Section 2 introduces literature review for teaching and learning Mathematics by using English language and Mobile Learning. Section 3 analyzes the factors affecting the learning of Mathematics by using the English language and the applications of this tool. Section 4 offers conclusions and development directions.

## II. LITERATURE REVIEW

### A. Teaching and learning Mathematics by using English language

Teaching and learning Mathematics by using the English language is increasingly popular and interesting in Vietnam. The reason is quite simple because learning Mathematics by using the English language not only helps students learn Mathematics, but also improves their English. Thus, it will provide a realistic environment for students to initialize using English in basic Science subjects, then continue to gain the capacity to integrate into global Education.

From the 2015 – 2016 school year, the Ministry of Education and Training has focused on implementing a policy to encourage schools from Primary to High School to teach Mathematics and Natural Sciences subjects by using English language and many parents agree to this suggestion. Nhat (2013) pointed out that teaching Mathematics by using English not only to the goal of teaching Mathematics but also aims to improve the English proficiency of students and teachers, which is the key for students to gain deeper expertise, consistent with the general trend of education in the world. Learning Mathematics thinking in English not only helps students develop intelligence but also develops their language skills, helping them to think critically and comprehensively.

Besides, Mathematics competitions in the English language are held with an increasing number in Vietnam and the world. In Vietnam, the most popular is the ViOlympic Mathematics contest, which is organized in English by the Ministry of Education and Training. Vietnamese students also participate in numerous Mathematics competitions outside Vietnam, such as Australian Mathematics Competition (AMC), International Mathematics Olympiad (IMO), Southeast Asian Mathematical Olympiad (SEAMO), International Mathematics Tournament of the Towns (ITOT), International Kangaroo Mathematics Contest (IKMC), and among others. With this trend, it is essential and significant for Students of Mathematics Teacher Education to have specialized English knowledge to study and teach in the near future.

### B. Mobile Learning

M-learning (mobile learning) is a form of learning that first appeared in the 90s of the twentieth century in universities in Europe and Asia. Currently, there are a lot of definitions of m-learning, based on the Cambridge dictionaries define m-learning is "learning done on electronic devices such as smartphones, laptops, computers and tablets". Thao (2013) indicated that m-learning is a form of training that takes place anytime and anywhere, regardless of the location of study and meets the highly personalized needs of learners, it consists of two inseparable elements, the use of technology devices and the learner's mobility. That is, regardless of where the learner, just owning a mobile device can participate in learning.

From these benefits, learning through smartphones is gradually becoming more popular and prevalent, especially foreign language learning, and it has also become a trend for the educational app market. According to Kim & Kwon (2012) statistics about foreign language learning applications searched from the iOS AppStore, the main focus of these applications is vocabulary: with 55% of the activities for learning vocabulary and vocabulary learning applications

account for 41%, other focus areas such as spelling and pronunciation are also closely related to vocabulary learning.

## III. RESEARCH METHODS

### A. Survey and analysis of specialized English learning of Students of Mathematics Teacher Education at Can Tho University

In order to find out the factors affecting English language proficiency in Mathematics of Students of Mathematics Teacher Education at Can Tho University. With the aim to build a direction to support students to learn Mathematics effectively in English. The research team created a survey on Google Form, then sent the online survey to graduate and master Students of Mathematics Teacher Education at Can Tho University. This study obtained 127 different surveys and full of information.

#### 1) Overview of analysis

##### a) Identify the initial factors that affect the level of confidence in specialized English

After consulting some Students of Mathematics Teacher Education at Can Tho University, this study identifies the initial factors that can affect students' level of confidence in specialized English as follows:

- Objective factor (NTKQ): Including course variables, this variable was converted into the number of years starting school (KH), gender (GT), orienting the implementation after graduation (TL).
- Subjective factor (NTCQ): Including English certificate variables (CC), the level of search specialized English terminology (TC), the level of using specialized English terminology (SD). The results of the level of confidence in specialized English (AG) are averaged by the confidence level of 4 skills: listening (LI), speaking (SP), reading (RE) and writing (WR).

#### b) Sample structure

The study obtained 127 surveys with the following statistical information:

TABLE I. SAMPLE STRUCTURE BY EACH OF THE OBJECT GROUP

#	Variable	Scale	Symbol	Quantity	Proportion
1	Course	1: Course 45	KH	9	7,14%
		2: Course 44		18	14,29%
		3: Course 43		15	11,90%
		4: Course 42		41	32,54%
		5: Course 41		20	15,87%
		6: Course 40		23	18,25%
2	Gender	1: Male	GT	59	46,83%
		2: Female		67	53,17%
3	Future orientation	1: Working outside the industry	TL	5	3,97%
		2: Teaching		82	65,08%
		3: Continue to study master degree		39	30,95%
4	English certificate	1: No certificate foreign language level B or higher	CC	96	76,19%
		2: Have certificate foreign language level B or higher		30	23,80%
5	The level of search Maths terminology	1: Never	TC	5	3,97%
		2: Seldom		3	2,38%
		3: Sometimes		12	9,52%
		4: Often		69	54,76%
		5: Very often		37	29,37%
6	Level of using Maths terminology	1: Never	SD	6	4,76%
		2: Seldom		40	31,75%
		3: Sometimes		34	26,98%
		4: Often		44	34,92%
		5: Very often		2	1,59%

It should be noted that the number of samples accounted for more than two-thirds of Students of Mathematics Teacher Education currently attending the school.

### c) Methods of data analysis.

The paper uses univariate and multivariate data analysis by analyzing the single and multivariate variance (ANOVA and MANOVA) to clarify the research problem. We conduct correlation and regression analysis: determine the linear correlation between AG and related factors. Finally, we build a logistic regression model to find the relationship between the classification of each learning outcome and the influencing factors.

### d) Methods of implementation.

The collected data will be encoded and entered into SPSS statistical software version 26 for processing. The analysis will take the following steps: First, analyze the impact of the course on the confidence level of four specialized English listening skills (AG). Next, analyze the influence of factors on AG: Identify each factor with statistical significance affecting AG. Finally, analyze the influence of factors on AG through logistic regression models. It should be remarked that the analysis is conducted at the 5% significance level.

### 2) Statistical analysis

From the survey results, we have the correlation coefficient table between the independent variables as follows:

TABLE II. THE PAIR CORRELATION COEFFICIENT OF VARIABLES INCLUDED IN THE MODEL

		KH	GT	TL	CC	TC	SD
KH	Pearson Correlation	1	-.100	.032	.515**	.258**	.262**
	Sig. (2-tailed)		.264	.726	.000	.004	.003
	N	126	126	126	126	126	126
GT	Pearson Correlation	-.100	1	-.086	.002	.035	-.025
	Sig. (2-tailed)	.264		.337	.984	.695	.779
	N	126	126	126	126	126	126
TL	Pearson Correlation	.032	-.086	1	.143	-.036	.001
	Sig. (2-tailed)	.726	.337		.109	.690	.993
	N	126	126	126	126	126	126
CC	Pearson Correlation	.515**	.002	.143	1	.174	.202*
	Sig. (2-tailed)	.000	.984	.109		.051	.023
	N	126	126	126	126	126	126
TC	Pearson Correlation	.258**	.035	-.036	.174	1	.536**
	Sig. (2-tailed)	.004	.695	.690	.051		.000
	N	126	126	126	126	126	126
SD	Pearson Correlation	.262**	-.025	.001	.202*	.536**	1
	Sig. (2-tailed)	.003	.779	.993	.023	.000	
	N	126	126	126	126	126	126

\*\* . Correlation is significant at the 0.01 level (2-tailed).

\* . Correlation is significant at the 0.05 level (2-tailed).

As seen from Table 2 that the variables TC and SD; KH and CC are closely related to each other, so in the

implementation of the classification problem we omit the variables TC and KH. Conducting logistic regression analysis, only the variables SD and TL are statistically significant at the 5% level when included in the model, the remaining variables are not significant at this level. This suggests that the variables do not play an important role in the level of confidence in specialized English. After removing variables with no statistical significance, we get the following table:

TABLE III. REGRESSION ANALYSIS RESULTS OF 2 VARIABLES

Coefficients							
	Unstand- ardized Coeffi- cients		Stand- ardized Coeffi- cients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Con- stant)	1,262	,240		5,255	,000		
TL	,265	,109	,185	2,439	,016	1,000	1,000
SD	,426	,064	,505	6,652	,000	1,000	1,000

It should be noted that Table 3 used to test hypotheses and Sig values. of TL, SD, i.e. futureoriented factors and use of specialized English terminology < 5% demonstrated that these two factors have a statistically significant impact on the average confidence level of 4 English skills. That means 4 assumptions are accepted.

In addition, looking at the coefficient B, it can be explained as follows, the coefficient B of the level of using the term is 0.426, meaning that when the variable level of using the term increases by 1 unit, the average value of confidence level of 4 skills increased by 0.426 units. Moreover, it is possible to compare and determine the level of influence of the factors: the greater the factor B, the more likely it is to comment that the factor has a higher level of influence than other factors in the research model. The Sig. value of the variables: English certificate, the level of search terminology and gender correspond to greater than 5%, these variables have no effect on the dependent variable. From the table above we have the following linear regression results:

$$AG = 0,265.TL + 0,426.SD + 1,262 (1)$$

where AG: average value of confidence level 4 English skills; TL: values average future direction (1: working outside the industry, 2: teaching, 3: continue to study master degree); SD: level of using specialized English terminology. This shows that the main factor affecting specialized English proficiency is the level of use of specialized English terms, the more you use it, the better your English will be. This is also the premise for the topic of designing Mathematical learning applications in English.

### B. Design Applications “English for Teaching and Learning Mathematics” (ETLM)

In order to promote low-level college students’ Mathematics Terminology learning motivation, the research team designs an application which introduces Mathematics Terminologies in both English and Vietnamese. Also, there is an example sentence followed by each word. A total of 32 units are designed, and there are about 550 words in total.

Apple iOS operating system was used in the app design in this research. The App prototype on Xcode 11.3 is provided in Fig. 1.



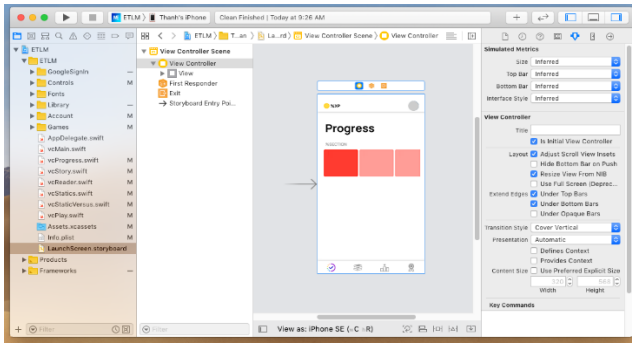
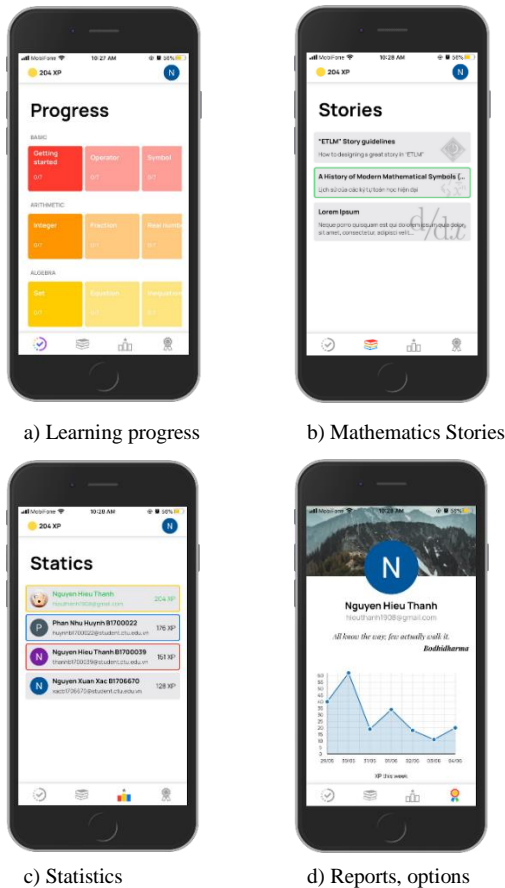


Fig. 1. App prototype on Xcode 11.3

By using the iOS software development kit and Swift programming language, the authors can customize the user interface of this application easily, which are shown in Fig. 2. After that, the researcher worked on the content and built the units in each grade. The final step is integrating data to the app.

The primary functions are constructed in the Tab bar (bottom of the phone-screen) such as Learning progress (Figure 2a), Mathematics stories (Figure 2b), Statics (Figure 2c), Reports and options (Figure 2d).



a) Learning progress

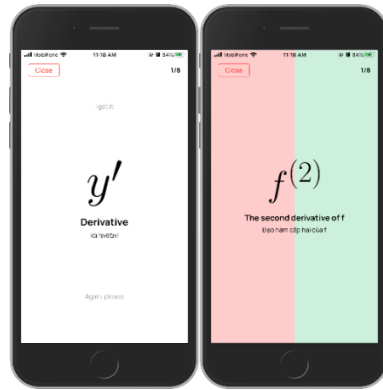
b) Mathematics Stories

c) Statics

d) Reports, options

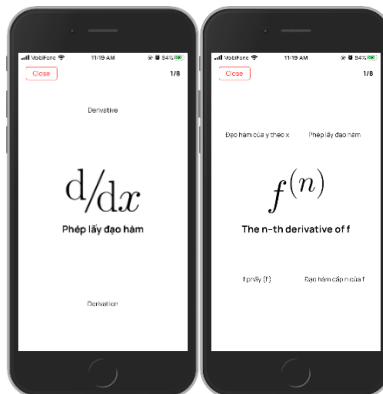
Fig. 2. The main functions of the application

To facilitate the distribution of applications across multiple platforms, we decided to host the lessons at a web-hosting, build API (Application Program Interface) for easier managing and accessing remotely. Figure 3 provided the interface of the application.

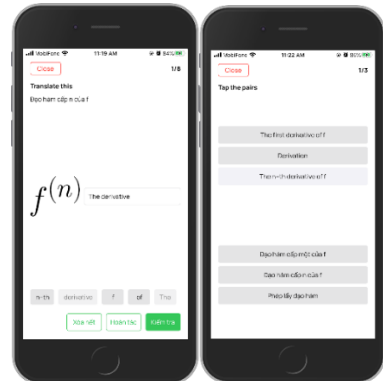


a) Learn word

b) Choose True/False

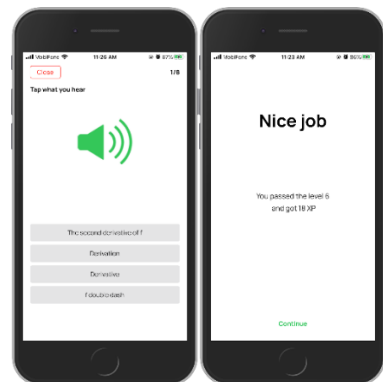


c) Select words or word's means



d) Sort words

e) Match words



f) Listening

g) Scores

Fig. 3. The interface of the application

Exercise-types used in the application include:

1. Learning word (Figure 3a): drag the image up to skip a word, down to repeat, users should repeat until you've memorized the word.
2. Choosing True/False (Figure 3b): Double-check the learners' knowledge when making random English words and meanings, drag pictures to the left (red areas) if they do not match, and vice versa, drag to the right (green areas).
3. Selecting the meaning of words (Figure 3c): Users need to drag and drop images into words/meanings that match it.
4. Sorting words into complete sentences (Figure 3d): Users select words/characters one by one, the sorted results are displayed in the middle frame of the screen, then select "Check" to confirm.
5. Matching words/meanings (Figure 3e): the player matches the word with the corresponding meaning (there are 3 pairs of words) by clicking on the word in one group and then selecting the corresponding meaning in the other group.
6. Listening (Figure 3f): Users can listen to a short audio clip, then select the audible option.
7. Scores (learning experience points, or experience points) are announced after completing a level (Figure 3g), which XP Scores = Standard XP + Bonus XP.

where: Standard XP is 10XP, the highest Bonus XP is 10XP is calculated as follows:

Bonus EXP = [(Number of words – Number of false answers)/Number of words × 10]

As in an illustration, a lesson has 15 vocabularies, every time you answer incorrectly, the learner will lose 1/15 of the total reward points, it is remarked that the error is more than 15 times, there will be no reward.

Next, the researcher designed the function to read Mathematics stories (Figure 2b) with bilingual content, and with illustrations. The main purpose is to help learners become familiar with reading and translating specialized documents when they have learned some specialized terms, in the reader interface, learners can toggle the translated version of a paragraph by tapping on the paragraph (Figure 4).



Fig. 4. The reading interface



Fig. 5. Learning process of other learners

Finally, the learning ranking function based on XP scores (Figure 2c) helps learners view details and compare skills scores with other learners by tapping on that person's name (Figure 5).

### C. App reviews

This paper has built the application ETLM on the iOS platform, with the aim of helping students and high school students learn Mathematics by using English proactively and easily control the learning process. With the data set in the beta-version, the application has been used by 2 classes of 2 teachers and 60 students to teach and learn bilingual mathematics. After 4 weeks of time, the research team created a meeting to interview them to evaluate the application. Overall, the application has received the majority of positive feedback from users about some criteria such as content, user interface and effectiveness. In other words, the application also has some advantages and disadvantages as follows:

In terms of advantages, this is a tool to support learning and teaching bilingual Mathematics in High School. With the support of the internet, you can learn Mathematics by using English anywhere and anytime. Furthermore, the visual image has several interactions so that the learning becomes easy and effective. In addition, this application also has a ranking function as a driving force to exceed and rank high. Besides, Mathematics stories make learners feel interesting and motivate them to learn more about English vocabulary, Mathematicians and Mathematical problems mentioned in that story.

Every coin has two sides, and it's true with this, that is, the application also has some limitations that users of this application need an internet connection and this application is not able to evaluate the pronunciation of learners.

With these reasons, we believe this is a very practical and meaningful tool for teachers and students in High School. We next compare with some popular English learning applications based on the research of Gangaianaran and Pasupathi (2017), the criteria given by Jareño et al. (2016):

TABLE IV. COMPARISON OF ETLM WITH POPULAR ENGLISH LEARNING APPS

Criteria	ETLM	Duolingo	Rosetta stone	Memrise
<b>Education criteria</b>				
Objects	High school students or above	High school students or above	High school students or above	High school students or above
Teaching skills	Vocabulary, grammar, reading comprehension, listening	Vocabulary, grammar, listening, pronunciation	Vocabulary, grammar, reading, listening, pronunciation	Vocabulary, grammar, listening, pronunciation
Guidelines	On website	On website	In-app guidelines	In-app guidelines
Easing of learning	Ascending	Ascending	Ascending	Ascending
Login methods	Google account	App account or Google account	App account or Google account	App account or Google account
<b>Technical criteria</b>				
Internet requirement	Yes	Free version	Free version	Free version
Intuitive interface	Yes	Yes	Yes	Yes
Advertising	No	Free version	Free version	Free version
App store rating	Not available	4.6/5 stars	4.8/5 stars	4.8/5 stars

#### IV. CONCLUSION

This article has designed an application to support the study of Mathematics terminology by using English through mobile platforms, thereby both approaching the m-learning trend and helping to improve Mathematics learning by using English among students. The results and feedbacks from the test user group show that this application has initially been successful and meaningful and it is expected to be a convenient learning tool for specialized students. Moreover, this application is provided free of charge and is regularly updated. Hence learners can easily evaluate and comment through the TestFlight offered by Apple, Inc.

Nevertheless, the team also encountered some difficulties during the implementation process, specifically to publish this application on the AppStore, it is required that a minimum of registration for participation in the Apple Developer Program, with the fee is 99 US Dollars, and need to own a device running macOS (MacBook, iMac, MacPro, etc.). But this equipment is often very expensive. Aims to be able to develop applications on the iOS platform, the project will continue to improve and expand the application's data set and publish to the AppStore.

#### REFERENCES

- [1] Ambarini, R., Setyaji, A., & Suneki, S. (2018). Teaching Mathematics Bilingually for Kindergarten Students with Teaching Aids Based on Local Wisdom. *English Language Teaching*, 11(3), 8-17.
- [2] Berger, A., & Klímová, B. (2018). Mobile application for the teaching of English. *Advanced Multimedia and Ubiquitous Engineering*. MUE 2018, FutureTech 2018. 518, 1-6. Springer, Singapore (2019).
- [3] Bione, T., & Cardoso, W. (2020). Synthetic voices in the foreign language context. *Language Learning & Technology*, 24(1), 169-186.
- [4] Clark, R. C., & Mayer, R. E. (2016). E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. John Wiley & Sons.
- [5] Dung, L. V. (2011). Strengthen the motivation of specialized foreign language learning for students of universities and colleges. *Journal of Language and Life*, (12), 6-10.
- [6] Dong, L. Q. (2011). English majors-some problems on teaching content. *Journal of Language and Life*, (11), 27-32.
- [7] Gangaianmaran, R., & Pasupathi, M. (2017). Review on use of mobile apps for language learning. *International Journal of Applied Engineering Research*, 12(21), 11242-11251.
- [8] Georgiev, T., Georgieva, E., & Smrikarov, A. (2004, June). M-learning-a New Stage of E-Learning. In *International conference on computer systems and technologies-CompSysTech*. 4(28), 1-4.
- [9] Hau, N. H., Tuan, B. A., Thao, T. T. T. & Wong, W.K. (2019). Teaching Mathematics by practical decision modeling in Vietnam High schools to serve the fourth industrial revolution. *Journal of Management Information and Decision Sciences*, 22(4), 444-461.
- [10] Heift, T. (2003). Drag or type, but don't click: A study on the effectiveness of different CALL exercise types. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée*, 6(1), 69-85.
- [11] Heift, T., & Nicholson, D. (2000, June). Theoretical and practical considerations for web-based intelligent language tutoring systems. In *International Conference on Intelligent Tutoring Systems* (pp. 354-362). Springer, Berlin, Heidelberg.
- [12] Jareño, A., Morales-Morgado, E. M., & Martínez, F. (2016, November). Design and validation of an instrument to evaluate educational apps and creation of a digital repository. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*. 611-618.
- [13] Kearney, M., Schuck, S., Burden, K., & Aubusson, P. (2012). Viewing mobile learning from a pedagogical perspective. *Alt-J-Research In Learning Technology*, 20(1).
- [14] Kim, H., & Kwon, Y. (2012). Exploring smartphone applications for effective mobile-assisted language learning. *Multimedia-Assisted Language Learning*, 15(1), 31-57.
- [15] Klimova, B. (2019). Mobile Learning and Its Impact on Learning English Vocabulary. *Advanced Multimedia and Ubiquitous Engineering*. MUE 2019, FutureTech 2019. 590, 271-276. Springer, Singapore.
- [16] Makoe, M., & Shandu, T. (2018). Developing a Mobile App for Learning English Vocabulary in an Open Distance Learning Context. *International Review of Research in Open and Distributed Learning*, 19(4).
- [17] Martin, I. A. (2020). Pronunciation development and instruction in distance language learning. *Language Learning & Technology*, 24(1), 86-106.
- [18] Mayer, R., Stull, A., Almeroth, K., Bimber, B., Chun, D., Bulger, M., ... & Zhang, H. (2009). Using Technology-Based Methods to Foster Learning in Large Lecture Classes: Evidence for the Pedagogic Value of Clickers.
- [19] Muhammed, A. A. (2014). The impact of mobiles on language learning on the part of English foreign language (EFL) university students. *Procedia-Social and Behavioral Sciences*, 136(9), 104-108.
- [20] Nhat, T. N. M. (2013). Teaching Mathematics and science subjects in English: Opportunities and challenges. *Journal of Language and Life*, 5, 17-22.
- [21] Sarraf, M., Elgamel, L., & Aldabbas, H. (2012). Mobile learning (m-learning) and educational environments. *International journal of distributed and parallel systems*, 3(4), 31.
- [22] Teodorescu, A. (2015). Mobile learning and its impact on business English learning. *Procedia-Social and Behavioral Sciences*, 180, 1535-1540.
- [23] Thao, T. T. P. (2013). Exploiting M-learning in credit training. *Journal of Science & Technology*, (12), 45 - 50.
- [24] Tuan, B. A., Pho, K. H., Huy, L. M., & Wong, W. K. (2019). STEMTech model in ASEAN universities: An empirical research at Can Tho University. *Journal of Management Information and Decision Sciences*, 22(2), 107-127.
- [25] Tuan, B. A., Huy, L. M., Anh, N. T. M., Tuan, T. Q., & Linh, P. V. Đ. (2019). Surveying computing programming thinking of students in STEM educational model: Research at Can Tho University. *Journal of Science Can Tho University*, 55, 56-61.
- [26] Wang, B. T. (2017). Designing mobile apps for English vocabulary learning. *International Journal of Information and Education Technology*, 7(4), 279.

# Fire Resistance Evaluation of Reinforced Concrete Structures

Dao Duy Kien

Faculty of Civil Engineering  
Ho Chi Minh City University  
of Technology and  
Education (HCMUTE)  
Ho Chi Minh City, Vietnam  
kiendd@hcmute.edu.vn

Do Van Trinh

Faculty of Civil Engineering  
Ho Chi Minh City University  
of Technology (HUTECH)  
Ho Chi Minh City, Vietnam  
dv.trinh@hutech.edu.vn

Khong Trong Toan

Faculty of Civil Engineering  
Ho Chi Minh City University  
of Technology (HUTECH)  
Ho Chi Minh City, Vietnam  
kt.toan@hutech.edu.vn

Le Ba Danh

Department of Bridge and  
Tunnel Engineering  
National University of Civil  
Engineering  
Ho Chi Minh City, Vietnam  
danhlb@nuce.edu.vn

**Abstract**—Alternative to fire tests in the near future may become modeling of structures under fire by means of the computer-aided design. With an objective source of data, the quantity of information from the created model is much higher than the results of standard fire tests. Based on the range of research a technique for modeling the precast concrete slabs to assess their fire rates with CAD system ANSYS has been developed. The adequacy of the simulation environment is confirmed by comparison with experimental data.

**Keywords**—Fire resistance, reinforces concrete, structures, modeling, finite element.

## I. INTRODUCTION

One of the most important stages of the building and structure design is a set of measures required to ensure the level of security over the estimated period of use, as well as in the emergencies.

For example, the structures should maintain their bearing and protecting capacity for the rated time necessary to ensure the safety of people, to protect assets and to provide liquidation and rescue. Understanding the basics of behavior of reinforced concrete structures under load, thermal effects and the impact of the environment allows performing efficient and safe design under the given criteria.

To date, there are several ways to get the desired results [1-5]. However, their use is tedious, time consuming and costly.

Alternative to existing methods of determining the fire resistance of building structures is their fire behavior modeling by means of computer-aided design. The use of specialized computer codes can significantly increase the profitability of the project works and increase their efficiency. Adequate modeling environment and objective source data can help to create a model of the building structure in the shortest possible time and with minimal cost of funds. The quantity of information will be much higher than at the results of standard fire tests.

## II. GENERAL REGULATIONS

The aim of work was to study the possibilities of modern software tools to predict the behavior and optimize the design of reinforced concrete used in buildings.

As the design environment for implementing the tasks, software modules of the finite element method (FEM) ANSYS were used. Testing of models was implemented in software applications, ANSYS Workbench and ANSYS

Mechanical. Choice was motivated by the need for verification of the solutions. It was implemented by two known methods – through a graphical interface and Workbench programming language APDL module Mechanical.

As the main evaluation scheme for theoretical model testing two types of floor plates: P-1 and P-2 [3] were chosen, both types are constructed with class C15/20 concrete (light expanded clay lightweight concrete) reinforced with longitudinal and transverse reinforcement [3]. The impact of 180 minutes' heat and uniformly distributed load  $q$  were taken into account as external factors. Baseline parameters of plates and load data are presented in Table I. Conventional load distribution corresponded to uniformly loaded beam on two supports. Heating was uniform over the entire length of the bottom surface (by reinforcement) too. Diagram is shown in Fig. 1.

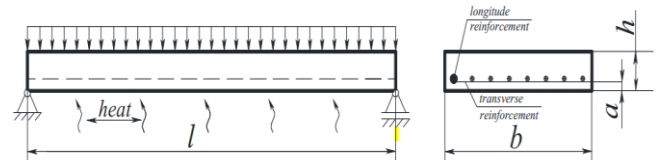


Fig. 1. Design scheme of plates (P-1, P-2)

TABLE I. CHARACTERISTICS OF PLATES

Parameters	Plates	
	P-1	P-2
Longitudinal reinforcement:		
- Number of bars.	8	8
- Diameter, mm	10	20
- Spacing between bars, mm	133.3	133.3
Transverse reinforcement:		
- Number of bars.	25	
- Diameter, mm;	6	
- Spacing between bars, mm	187.5	
Distance from the bottom edge to center of reinforcement (a), mm	25	30
Plate thickness (h), mm	120	
Width of the plate (b), mm	1200	
Plate length (l), mm	3000	
Distributed load (q), kN/m	10.59	22.95

The analysis was performed on the results of the solution of two types of problems in the following sequences [6,7]: thermodynamics analysis – the result of the temperature field distribution in terms of plate as a function of time; strength analysis – the result of the strain and stress distribution in terms of plate as a function of time and temperature. This included the possibility of plastic deformation of materials and concrete failure due to the cracks formation [8,9].

In this formulation the thermodynamics analysis was the primary one, since its results were used in the form of raw

data for strength analysis. Mathematical approaches that described the stress-strain state and the temperature field distribution of the model were significantly different, therefore, for each of them different types of FE were used [10].

The experimental curves of the concrete slab heating over the section were used for verification of the model by comparing the thermal analysis results. Experimental data on the temperature field distribution of the selected size of plates are known from the reference [3]. These comparisons are shown in Tables II, III and Fig. 2, 3, 4.

TABLE II. THE TEMPERATURE DISTRIBUTION OVER THE CROSS SECTION OF P-2 PLATE

Position of the layer	Source	Temperature layers at time t, min								
		20	40	60	80	100	120	140	160	180
Heating surface	experiment (in camera)	781	885	945	990	1025	1049	1072	1092	1110
	calculation	781	885	945	990	1024	1050	1070	1090	1109
Reinforcement	experiment	250	400	500	570	630	660	700	730	770
	calculation	192	310	420	540	630	652	715	740	782
Cooled surface	experiment	20	45	60	70	80	85	100	120	140
	calculation	20	40	48	52	67	88	109	127	144

TABLE III. THE TEMPERATURE DISTRIBUTION OVER THE CROSS SECTION OF P-1 PLATE

Position of the layer	Source	Temperature layers at time t, min								
		20	40	60	80	100	120	140	160	180
Heating surface	experiment (in camera)	781	885	945	990	1025	1049	1072	1092	1110
	calculation	781	885	945	988	1025	1049	1073	1092	1110
Reinforcement	experiment	250	400	500	570	630	660	700	730	770
	calculation	230	400	525	610	680	710	760	800	845
Cooled surface	experiment	20	45	60	70	80	85	100	120	140
	calculation	20.1	42	51	60	71	93	114	133	149



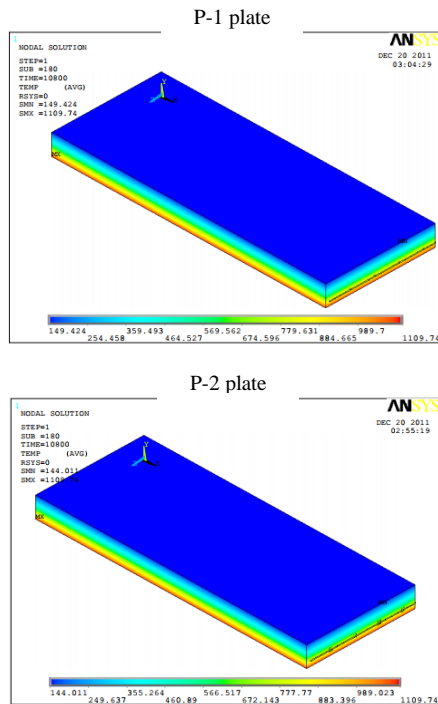


Fig. 2. The temperature distribution in the volume of plate

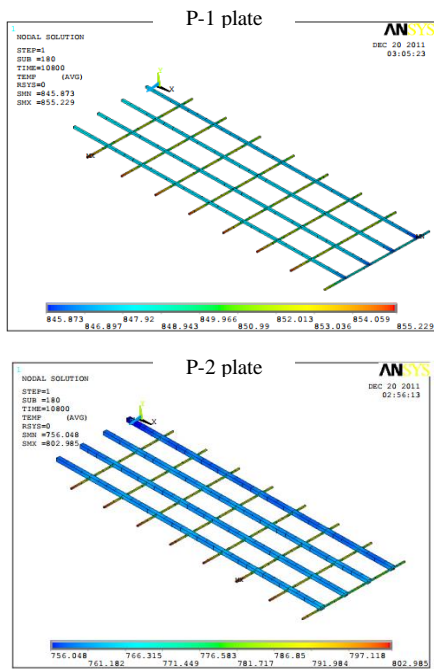


Fig. 3. The temperature distribution in the amount of reinforcement

As follows from the calculation, the accuracy of temperature fields data obtained using the finite element method is quite high. There is some error, especially for plate P-1 in the temperature of the reinforcement. This desynchronization can be caused by an error of the experimental data. According to the source [3], the experimental temperature in the reinforcement was determined by measuring the layer of concrete on the longitudinal reinforcement center depth.

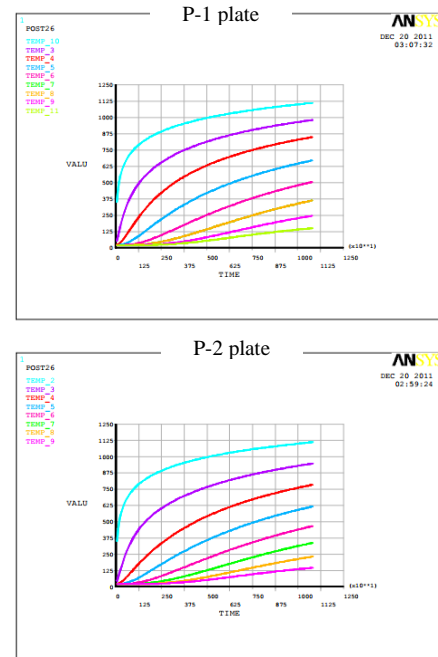


Fig. 4. The temperature versus time in the layers of plates

The differences in the temperature of the cooled surface are small and may not have a significant effect on the level of physical and mechanical properties of concrete. The obtained results demonstrate high accuracy of the finite element analysis.

In view of the requirements for the convergence of the analysis the concrete cracking and crushing in compression were not performed. In connection with mentioned the evaluation was carried out on the fire endurance level equivalent stresses in the reinforcement and magnitude of ultimate strains in materials.

In view of the requirements for the convergence of the analysis the concrete cracking and crushing in compression were not performed. In connection with mentioned the evaluation was carried out on the fire endurance level equivalent stresses in the reinforcement and magnitude of ultimate strains in materials.

- The equivalent von Mises stress;
- The equivalent von Mises stress;
- Stress profile height plate.

Strain state of the plate is determined by the following factors:

- Full deformation (mechanical and thermal);
- plastic deformation;
- Temperature deformation.

For each indicator, output is displayed on the following criteria:

- Deformations of the von Mises;
- Longitudinal strain (along the axis X);
- Strain profile height plate.

The analysis of the stress-strain state in comparison with experimental data as a function of temperature is presented in tables IV and V.

TABLE IV. EXPERIMENTAL AND THEORETICAL CHARACTERISTICS OF THE STRESS – STRAIN STATE OF PLATE P-1 EXPOSED TO HEATING AND LOADING

Parameter	Data source	Heating time, min								
		20	40	60	80	100	120	140	160	180
The equivalent stress in the reinforcement, MPa	experiment	226.1	231.2	189.3	173.0	196.6	-	-	-	-
	calculation	212.0	200.0	195.0	182.0	175.0	-	-	-	-
Compressive stress in concrete	experiment	8.48	9.22	9.60	9.93	10.31	-	-	-	-
	calculation	10.0	12.5	12.8	13.1	13.5	-	-	-	-
Strains of the concrete, %	experiment	0.08	0.10	0.11	0.14	0.20	-	-	-	-
	calculation	0.089	0.12	0.125	0.15	0.23	-	-	-	-
Strains of the reinforcement, %	experiment	0.30	0.50	0.60	0.81	1.32	-	-	-	-
	calculation	0.30	0.54	0.65	0.87	1.38	-	-	-	-
The maximum deflection, mm	experiment	43.0	61.0	68.0	89.0	137.0	-	-	-	-
	calculation	39.0	64.0	72.0	104.0	160.0	-	-	-	-

TABLE V. EXPERIMENTAL AND THEORETICAL CHARACTERISTICS OF THE STRESS – STRAIN STATE OF PLATE P-2 EXPOSED TO HEATING AND LOADING

Parameter	Data source	Heating time, min								
		20	40	60	80	100	120	140	160	180
The equivalent stress in the reinforcement, MPa	experiment	132.0	139.0	134.0	127.0	125.0	119.0	115.0	113.0	117.0
	calculation	125.0	130.0	132.0	120.0	120.0	115.0	113.0	110.0	110.0
Strains of the concrete, %	experiment	0.07	0.14	0.16	0.18	0.20	0.22	0.25	0.28	0.31
	calculation	0.08	0.16	0.17	0.20	0.21	0.25	0.30	0.32	0.34
Strains of the reinforcement, %	experiment	0.06	0.25	0.39	0.50	0.58	0.69	0.80	0.97	1.20
	calculation	0.08	0.32	0.42	0.52	0.60	0.76	0.87	1.05	1.30
The maximum deflection, mm	experiment	14.0	40.0	56.0	68.0	80.0	93.0	107.0	124.0	160.0
	calculation	28.0	52.0	68.0	78.0	89.0	104.0	115.0	142.0	186.0

Certain differences in design stresses are caused by deflection sensitivity of the model to the values of linear expansion coefficients with a significant temperature gradient through the thickness of the slab. In the reference [1-3] for a concrete class C15/20 (light expanded clay lightweight concrete) given coefficient of linear thermal expansion (TCLE) materials differs significantly.

Temperature component in the deflection of the beam reaches 70-80% of the maximum, therefore, even minor fluctuations TCLE able to exercise significant influence on the overall strain state of the structure.

Overall, the obtained results are quite close to the experimental data, particularly in regard to the bearing deformation reinforcement. Due to advances in concrete yield strength at high temperatures, the bearing capacity of the reinforcement is the main factor for fire resistance of reinforced concrete slabs.

According to the thermo-mechanical deformation diagrams of reinforcing steel, early stages of deformation strength are 1,25% [10-12]. When reaching this value at temperature influence is ductile failure of the material, fixtures loses load capacity.

According to experimental data presented in table 4 and 5, the ductile failure reinforcement occurs at temperatures of 100 °C for plate P-1 and P 160-170 °C - for plate P-2, which agreed with the calculated results.

### III. CONCLUSIONS

The results of the research can be used in calculations of fire resistance of reinforced concrete frame building designs for computer-aided design.

The main results obtained in this paper:

- The analysis of the ANSYS computing environment components was developed. Their ability to record the temperature and force effects that arise in the construction of a fire was reviewed;
- The experimental data of reinforced concrete slabs in fire behavior were studied, including heat exposure of standard and real fire;
- A computational model of reinforced concrete slabs in-fire behavior was developed in order to assess their fire rates;

The method for fire resistance assessment of reinforced concrete slabs was developed, allowing the use of it in computer-aided design.

#### REFERENCES

- [1] A.F. Milovanov. Fire resistance of concrete structures. Moscow: Stroyizdat, 1998. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] Teaching aid for the design and construction companies. Section I: Fire protection of tall buildings and unique objects. Moscow: PKF “Endemic”, 2004.
- [3] E.V. Levitsky. Diagrammatic method for calculating the fire resistance of the static problem of reinforced concrete structures. PhD dissertation. Moscow: Russian State Library, 2007.
- [4] International Organization for Standardization. Fire-resistance tests. Elements of building construction. Part I. General requirements: ISO 834-1:1999(E). Geneva: ISO, 1999.
- [5] Agency Standard. Rules on the fire and post fire resistance of concrete structures: STO 36554501-006-2006. Moscow: FSRC “Construction”, 2006.
- [6] ANSYS Support Centre. Available: [http:// www.cae-expert.ru/](http://www.cae-expert.ru/) [Accessed: May 24, 20].
- [7] European standard. Eurocode 2. Design of concrete structures. Part 1-2. General rules. Structural fire design : EN 1992-1-2:2004(E). Brussels: CEN, 2004.
- [8] S.P. Timoshenko. Strength of materials. Moscow: Science, 1965.
- [9] V.A. Kudryashov. The basic heavy concrete stress-strain diagrams transformation for a short-term high-temperature impact. Institute for Command Engineers Bulletin, 2008, 1/7. Minsk.
- [10] V.A. Bruyako. Engineering analysis in ANSYS Workbench: Samara, 2010.

# Digital Economy: Overview of Definition and Measurement Criteria

Nguyen Thi Thanh Van

*Economics Faculty*

*Ho Chi Minh city University of Technology and Education*

Ho Chi Minh city, Vietnam

vanntt@hcmute.edu.vn

Nguyen Thien Duy

*Administration Office*

*University of Economics Ho Chi Minh city, Vietnam*

Ho Chi Minh city, Vietnam

thienduy@ueh.edu.vn

**Abstract**— Nowadays, the rapid development of information technology has opened up new development trends for countries. The concept of "Digital Economy" appears and is mentioned as an inevitable trend in the future. However, due to the advancement of technology platforms which results in such quick changes and applications in many fields that the concept of digital economy has not yet been defined accurately and universally, nor yet been clarified which activities are included in the measurement for the digital economy. The objective of this study is to give an overview of digital economy concepts and ways of measuring digital economy used by countries around the world. Further studies can be screened and adjusted to create a set of criteria for measuring digital economy for Vietnam

**Keywords**— *digital economy, measuring the digital economy, information technology*

## I. INTRODUCTION

The boom of the Industrial Revolution 4.0 brings about changes in the lives of global people. Particularly, in developing countries, this transformation takes place at a rapid pace and is associated with real life. People use services like e-commerce, e-banking, e-learning, etc. more often, and those services gradually become an important part of their lives. Businesses are also adapting quickly, they are interested in innovation, application of science, and technology in business activities. Technical phrases like internet of things, big data, cloud computing become familiar words. The government emphasizes the trend of digitizing the economy and social activities, considering it as an important contribution to GDP (Gross Domestic Products). From theoretical research, digital economy comes to life. However, digital economy is quite abstract and flexible in measurement. Therefore, this article aims to clarify the definition of digital economy, and gives an overview of some criteria for measuring digital economy which are used in the world and in Vietnam. The results of this review will pave the way for further studies on digital economy measurement..

## II. DEFINITION OF DIGITAL ECONOMY

The concept of digital economy first appeared in the mid-1990s and underwent many changes that reflected the change of technology [1].

In the late 1990s, the analysis was primarily concerned with Internet adoption and the term referred to as "Internet economy". At this time, only primitive ideas formed about the impact of the internet on the economy [2] and they only emerged in developed countries. Then, in the mid-2000s, when the Internet was widely used, reports began to focus on analyzing digital policies and technologies, on the other hand,

the development of information technology led to the digital orientation which became a core element of companies [3].

Over the past few years, countries and researchers have focused on spreading ways of digitizing services, products, and technologies across economies. This digitization process is considered as the transformation of businesses through the use of technologies and digital products [4]. If previously, digital was limited to several high-tech fields, nowadays high-tech products and services are making rapid changes on a large scale and in many areas.

Due to the rapid development of technology platforms, drastic changes and applications in many fields, the concept of digital economy has not been defined accurately and universally, nor has it been clarified what activities are included in the digital economy measurement. We cite some definitions of digital economy according to the change of technology as follows:

- Tapscott had no specific definition but the author explained that digital economy combines intelligence, knowledge, and creativity to make breakthroughs in achieving wealth and social development [2].
- Margherio et al. also had not given a specific definition, but this is considered the first study to address the key aspects of digital economy: Internet, e-commerce, delivery of goods and services via digital delivery, retail of tangible goods [5]

Recently, as the advent of Industrial Revolution 4.0 gets clearer, the definition of digital economy has also been given by many organizations:

- G20 - The Digital Economy Task Force (DETF) defined as "a broad range of activities that include the use of digitised information and knowledge as the key factor in production, of modern information networks as an important activity space, and of Information Communication Technology (ICT) as an important driver of productivity growth and economic structure optimisation" [6].
- Oxford University Press (OUP) defined as "digital economy as an economy which functions primarily by means of digital technology, especially electronic transactions made using the Internet" [7].
- According to the Bureau of Economic Analysis (BEA), the definition of Digital Economy is mainly based on the Internet, related information, and communication technology [8].

- In Vietnam's future digital economy toward 2030 and 2045, Cameron et al. used the concept of digital economy as "all businesses and services that have a business model based primarily on selling or servicing digital goods and services or their supporting equipment and infrastructure" [9].

Thus, although the definitions can vary, they all lead to a common point: economic activities based on the internet and information technology. Digital economy can be considered and calculated on the broadest and narrowest definitions as Fig. 1:

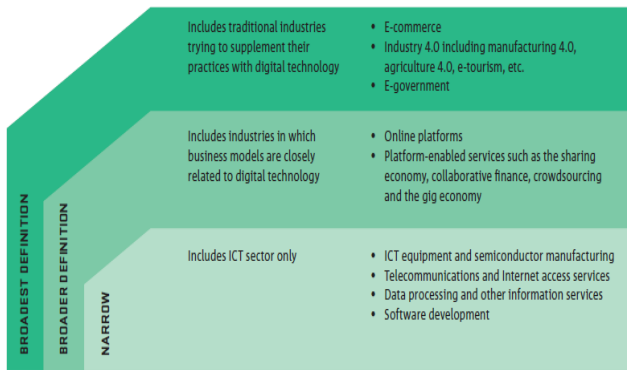


Fig.1. Broadest and narrowest definitions of the digital economy [9]

However, from this definition, a challenge arises in the way of measuring digital economy, because of differences in approach, in the arrangement of products into areas considered to be related to digital. Herrero and Xu also admitted that the measurement for digital economy is less clear, especially between countries [10]. For example, in China when compared with the Organisation for Economic Co-operation and Development (OECD), only some products in some areas have connections with the digital economy. This is because the product information is not consistent in all areas.

Variations in digital economy measurements will lead to significant differences when comparing this indicator between countries, as well as different views on the role and contribution of digital economy to the national economy. As such, it is essential to unify the measurement standards of the digital economy in the world.

Here, we will summarize some of the measurement methods that have been performed by organizations around the world

### III. MEASURING THE DIGITAL ECONOMY

#### A. United Nations Conference on Trade and Development (UNCTAD)

According to UNCTAD, digital economy can be divided into three main aspects [11]:

- Core or fundamental aspects of the digital economy, including fundamental innovation (semiconductors, processors), core technologies (computers, telecoms equipment), and infrastructure (internet and telecoms networks).
- Digital and information technology sectors, whose key products and services are based on core digital

technologies, including digital platforms, mobile applications, and payment services. This aspect affects many other sectors and makes a growing contribution to the economy.

- A broader aspect of digitization, including places where digital products and services are increasingly used (for example, e-commerce). As a result of digital technology, many sectors of the economy have followed the digitization path, leading to the emergence of many transforming businesses or business models such as finance and media, tourism, and transportation. Moreover, although less frequently mentioned, proficiency in digital skills of workers, consumers, and users is also important for the development of the digital economy.

UNCTAD uses the narrow definition as shown in Fig. 1 above, that is, only ICT sector is calculated, using the criteria as in Table I:

TABLE I. MEASUREMENT CRITERIA OF UNCTAD [11]

Components	Subcomponents
<b>ICT Manufacturing Industries</b>	
	Manufacture of computer, electronic & optical products
	Electronic components & boards
	Computers & peripheral equipment
	Communication Equipment
	Consumer electronics
	Magnetic & optical media
<b>ICT Trade Industries</b>	
	Wholesale of computers, computer peripheral equipment and software
	Wholesale of electronic and telecommunications equipment and parts
<b>ICT Services Industries</b>	
	Software publishing
	Telecommunications
	Computer programming, consultancy and related activities
	Information service activities
	Data processing, hosting and related activities; web portals
	Repair of computers and communication equipment
	Repair of computers and peripheral equipment
	Repair of communication equipment

#### B. Bureau of Economic Analysis (BEA) – U.S department of eCommerce

According to BEA in the concept of the digital economy, there are 3 aspects [8]:

- The digital infrastructure that allows computer networks to exist and operate: Computer hardware, computer software, telecommunications equipment and services, infrastructure, internet of things.
- E-commerce: Digital transactions take place via computer network system (e-commerce) in all 3 types of Business-to-Business (B2B), Business-to-Consumer (B2C), Peer-to-Peer (P2P).
- Digital media: users use products in paid or free digital format, and large data is collected and exploited by companies.



When measuring the digital economy, BEA is conducted through hardware, e-commerce, and digital media, support services, software, and telecommunications. Specifically with the following criteria as table II:

TABLE II. MEASUREMENT CRITERIA OF BEA [8]

Components	Subcomponents
<b>Infrastructure</b>	
	Computer hardware
	Software
	Telecommunications equipment and services
	Structures
	Internet of things
	Support services
<b>E-commerce</b>	
	Business-to-business (B2B) e-commerce
	Business-to-consumer (B2C) e-commerce
	Peer-to-peer eCommerce
<b>Digital media</b>	
	Direct-sale digital media
	“Free” digital media
	Big data

### C. European Unions (EU)

The European Union is oriented towards the digital economy and society. They have been measuring this concept through The Digital Economy and Society Index (DESI). This index has 5 main components as table III [12].

TABLE III. MEASUREMENT CRITERIA OF EU [12]

Components	Subcomponents
<b>Connectivity</b>	
	Fixed broadband
	Mobile broadband
	Fast broadband
	Ultrafast broadband
	Broadband price Index
<b>Human capital</b>	
	Internet user skills
	Advanced skills and development
<b>Use of internet</b>	
	Internet use
	Activities online
	Transactions
<b>Integration of Digital Technology</b>	
	Business digitisation
	e-Commerce
<b>Digital public services</b>	
	e-Government
	e-Health

### D. G20 – DETF (The Digital Economy Task Force)

DETF was established in 2017 to establish a toolkit to become the standard digital economy measurement for countries. DETF is based on existing indicators and therefore does not create new content but instead standardizes indicators, making it easier for countries to access and use for their own countries.

The G20-DETF toolkit is based on the broad definition in Fig. 1 and also on the sources of the OECD, the International Telecommunication Union (ITU), UNCTAD, EU, The World Bank Group (WBG), the International Monetary Fund (IMF), and the International Labor Organization (ILO). The introduced toolkit includes the following criteria as table IV:

TABLE IV. MEASUREMENT CRITERIA OF DETF [6]

Components	Subcomponents
<b>Infrastructure</b>	
	Investing in Broadband
	The rise of Mobile Broadband
	Toward higher Internet speed
	Prices for connectivity
	Infrastructure for the Internet of Things
	Secure servers infrastructure
	Household access to computers
	Household access to the Internet
<b>Empowering society</b>	
	Digital natives
	Narrowing the digital divide
	People's use of the Internet
	E-consumers
	Mobile Money
	Citizens interacting with government
	Education in the digital era
	Individuals with ICT skills
<b>Innovation and Technology Adoption</b>	
	Research in machine learning
	AI-related technologies
	Robotisation in manufacturing
	R&D in information industries
	Supporting business R&D
	ICT-related innovations
	ICT Use by businesses
	Cloud computing services
<b>Jobs and Growth</b>	
	Jobs in the information Industries
	Jobs in ICT occupations
	ICT workers by gender
	E-Commerce
	Value added in information industries
	The extended ICT footprint
	ICT Investment
	ICT and productivity growth
	ICT and Global Value Chains
	Trade and ICT Jobs
	ICT goods as a percentage of merchandise trade
	Telecommunications, computer, and information services as a percentage of services trade

### E. Vietnam

The digital economy is the trend of the world, and Vietnam is not an exception. The boom of digital technology presents many challenges and opportunities for Vietnam in particular and developing countries in general. Digital technology will particularly affect traditional industries, such as agriculture, tourism, and transportation in developing countries. The most important economic changes may also occur through the digitalization of traditional fields rather than through the emergence of new digital industries.

Over the past years, Vietnam has also identified the path of the economy, determined its position in the digital economy picture of the world and the region, as well as issued many policies to support the digital economy. However, as mentioned above, there's a lack of common standards for concept and measurement criteria for digital economy.

In Vietnam's future digital economy toward 2030 and 2045, Cameron et al. proposed the Digital Adoption Index (DAI). DAI is calculated from the data of companies which are the representatives in the manufacturing and agriculture sector of Vietnam. DAI's goal is to discover businesses in

today's digital application and their awareness of digital transformation. DAI is expected to be a good reference for government agencies to make appropriate investment policies and programs. DAI also helps businesses identify their position in the digital transformation journey, as well as analyze the capabilities, potentials and barriers to progress [9].

DAI consists of 6 criteria with many sub-criteria as follows:

TABLE V. DIGITAL ADOPTION INDEX (DAI) [9]

Components	Subcomponents
<b>Strategy and organisation</b>	Existence of digital strategy, digital roadmaps, etc.
	Leader support
	Existence of central coordination unit for digital adoption
	The suitability of the existing business model to digital adoption
	Regulation and suitability to technological standards and IP protection
<b>Finance</b>	
	Level of investment in digitalisation in the last year
	Level of investment in digitalisation in the next 3 years
<b>Infrastructure</b>	
	The level of infrastructure to support digital adoption (energy, telecommunication, transport, etc.)
	Connectivity quality
	The competence of the existing ICT system and requirements for digital adoption
	The level of cybersecurity methods in the business
<b>Human resources</b>	
	ICT skills of employees
	The extent that the business applies digital technologies to daily operations
	Training and retraining in digital related areas
	Business culture in terms of knowledge sharing, open innovation, etc.
<b>Smart production</b>	
	Application of advanced production management techniques (autonomous production line, FMS, CIM, etc.)
	Application of other digital technologies in production (blockchain, robotics, sensors, etc.)
	Level of digitalisation of production equipment
	The level of real-time data collection and utilisation
	The extent that the business has a real-time view on production
<b>Forward and backward linkages and logistics</b>	
	The amount the business uses multiple integrated sale channels
	The level of multiple information channel usage
	The level of automation and digital integration in logistics (from order capture, inventory management to warehousing and transportation)
	Collaboration among different players in the value chains
	Utilisation of customer data and consumer's digital competence

#### IV. CONCLUSION

The overview of the definition of the digital economy as well as its measurement criteria shows that there is a lack of uniformity between organizations or countries. Differences can come from an approach perspective: by narrow or broad definition; from determining the components of the digital economy or determining which goods and services link to the digital economy. All of these gaps require a common, agreed set of criteria for countries.

Vietnam in this period of rapid development also requires the calculation of the digital economy index. By 2025, the digital economy is aimed to account for 20% of GDP, which means that determination of a digital economy measurement criteria is extremely important. This opens up opportunities for the next study to be done to benchmark Vietnam's DAI against the criteria set by BEA, G20-DETF, or UNCTAD in order to select or supplement the criteria, therefore, to be completed and put in accountancy for Vietnam.

#### REFERENCES

- [1] R. Bukht and R. Heeks, "Defining, conceptualising and measuring the digital economy," GDI Development Informatics Working Papers, 2017, No. 68(0), pp.1-24.
- [2] D. Tapscott, *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*, McGraw-Hill, New York, NY, 1996.
- [3] OECD, *Measuring GDP in a Digitalised Economy*, OECD, Paris, 2016
- [4] S. Brennen and D. Kreiss, *Digitalization and digitization*, Culture Digitally, 2014.
- [5] L. Margherio, D. Henry, S. Cooke, and S. Montes, *The Emerging Digital Economy*, U.S. Department of Commerce, Washington. 1998.
- [6] G20 – DETF, *Toolkit for measuring the Digital Economy*, 2018.
- [7] OUP, *Digital Economy*, Oxford Dictionary, Oxford University Press, Oxford, UK. [https://en.oxforddictionaries.com/definition/digital\\_economy](https://en.oxforddictionaries.com/definition/digital_economy), 2017.
- [8] K. Barefoot, D. Curtis, W. Jolliff, J. R. Nicholson, R. Omohundro, *Defining and Measuring the Digital Economy*, Bureau of Economic Analysis, Washington, DC, 2018.
- [9] A. Cameron, T. Pham, J. Atherton, *Vietnam Today – first report of the Vietnam's Future Digital Economy Project*. CSIRO, Brisbane, 2018.
- [10] A. G. Herrero and J.Xu, "How big is China's Digital Economy," Working Papers, 2018, No. 04.
- [11] UNCTAD, *Digital Economy Report – Value Creation and Capture: Implication for Developing Countries*, United Nations Publications, 2019
- [12] DESI, *Digital Economy and Society Index*, European Commission, 2019.
- [13]

# Receptionist and Security Robot Using Face Recognition with Standardized Data Collecting Method

Quang-Minh Ky  
Faculty of Electrical  
and Electronics Engineering,  
HCMC University of  
Technology and Education,  
Ho Chi Minh City, Vietnam  
quangminh.910@gmail.com

Dung-Nhan Huynh  
Faculty of Electrical  
and Electronics Engineering,  
HCMC University of  
Technology and Education,  
Ho Chi Minh City, Vietnam  
16142024@student.hcmute.edu.vn

My-Ha Le  
Faculty of Electrical  
and Electronics Engineering,  
HCMC University of  
Technology and Education,  
Ho Chi Minh City, Vietnam  
halm@hcmute.edu.vn

**Abstract**—Face recognition has become the front runner for deep learning applications in the real world and this paper focuses on its implementation in a human-robot interaction and security system. For this specific project, it is inherent that restraints are created to allow the system to produce greater performance within the requirements of a receptionist and security robot. A k-nearest neighbors classifier is applied to further enhance the accuracy of face recognition. By sequencing images from videos, we create large datasets to train our own classifier in various conditions to increase its accuracy and lower false-positive rates in poor lighting environments. With the goal of creating a service robot, we have standardized our method of data collection for new inputs that will assist the recognition process in variable conditions of operation. The resulting product is a system that can accurately predict known and unknown faces with Asian features.

**Index Terms**—face recognition, human robot interaction, deep learning, standardized dataset, receptionist and security, lighting, images sequencing, k-nearest neighbor classifier

## I. INTRODUCTION

Face recognition has long since moved out of the realm of science fiction and became one of the foundations for advancements in neural networking, deep learning and artificial intelligent. While much work have been done [1]–[3], on the matter of enhancing older systems and even creating new ones altogether, the challenges of real-world applications still loom over the credibility and the potential of this technology as a modern solution to bio-metric security, identity verification, etc. Despite the amazing efforts that have been put forth by the computer vision community [4], flaws were found during our test with Asian faces in live feed videos and images shown in Figure 1, of which false positives were recorded with high frequency between different people. With the system being tested and used for customer services and security purposes in Vietnam, we proposed a solution to its implication in countries with distinct features and a method of collecting needed data to further increase the accuracy of recognition.

In this paper, our approach is to improve upon the face recognition aspect of the system as a k-nearest neighbor (kNN)

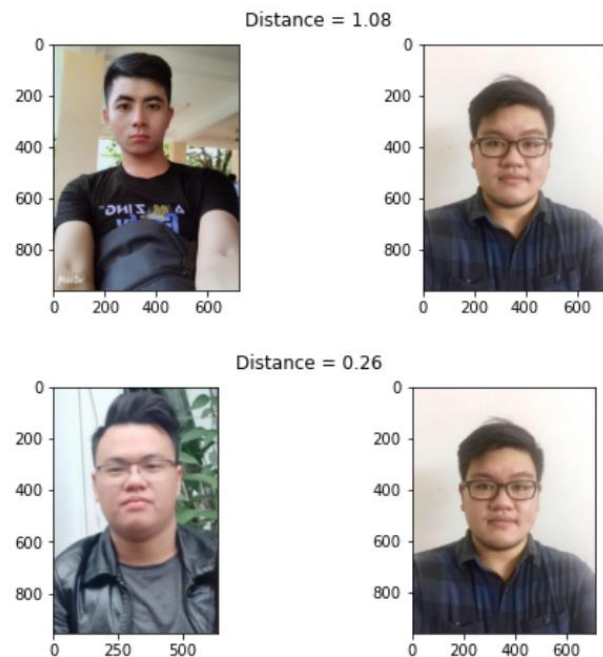


Figure 1. **Similarities between individuals.** This figure shows the Euclidean distance between two pairs of faces of three different people using FaceNet. Of which the closer the distance is to 0.0, the more alike they are and vice versa. If we were to have a threshold of 0.3, half of the standard testing threshold of 0.6 (the lower the threshold the more similar the faces have to be to be considered the same), the bottom pair would be considered the same person.

problem instead of trying to build a new model for embedding creation. As part of the solution, we added our own trained kNN classifier to the output of the deep convoluted neural network (DCNN) of FaceNet model. This is the simplest way to accurately identify and verify faces for small settings that a receptionist robot can be used in without exerting tremendous resources on training a DCNN model with large enough data to represent high accuracy and significantly low margin of error.

Furthermore, to allow the distance threshold to be as reasonably low as possible so that the inaccuracy of the model does not manifest upon the results and not to overfit the data, we devised a method of collecting facial information for new inputs in the kNN classifier. This will ensure that the amount of information on known faces is sustainable and the machine will be less susceptible to similarities between individual when a face is well defined. We will show the significant differences when the size of data vary in different environments.

The rest of the paper will follow the structure of: section II will present other works within the spectrum of this project; section III-A is a detailed explanation of our current detection model; section III-B emphasizes in the output of the face recognition model and section III-C is the application of kNN classifier into the recognition model; section IV describes the method of collecting facial data of a new face in a standardized way; section V is the showcasing of quantitative results of the robot; Lastly section VI, we will draw a definite conclusion based on the evidence gathered thus far while exploring the future prospects and means of improvements.

## II. RELATED WORKS

Prior to our work, the concept of a robot working in any spectrum of social interaction has been adopted worldwide and along with that, has also been subjected to many different ideologies. As the flagship for human-machine interactive robots, despite the inversely proportional relationship between exposures and acceptances [5], Japan has had an enormous amount of robotics invention that spreads throughout many different fields [6]. We not only try and improve the capabilities of a face recognition system in this paper alone but also explore the feasibility of such technology in the many aspects of real world application, one of which is the ability to be of use in customer services as a receptionist.

Gaining popularity in 2014 was Pepper the robot [7], a machine that can interact with people, read and respond to emotions, have spatial awareness, and mimic human-like gestures. This was not the first instance of socially interactive robot ideas, many approaches were studied [8], [9] and explained with great details the creative process of developing these machines. While these works have been focused on the mechanical aspects of a functioning receptionist robot, we believe that our project can work in unison with the preceding studies to create a more robust and adaptable system where the robot can remember and differentiate people and interact with more finest among different encounters.

Aside from the subject of human-robot interactions, flaws from face recognition technologies have dated even before the movement into IoT, and data science gained momentum. While Labeled Faces in the Wild (LFW) [10] has been the benchmark for mostly every face recognition architecture that was built, the diversity of the dataset, despite being classified by many models with accuracy upwards of 99%, these models were still found lacking when done in various lighting environment with multiple similar faces. There have been tremendous efforts put forth by the computer vision and data science community to

create datasets for model training, notably a 360K images of 2019 individuals with Asian faces [11]. The work done based on this massive amount of data has yet to give us many notable results due to our limitations. However, there is much to have faith in future endeavors in creating a DCNN model that can accurately differentiate every single individual.

While data is essential towards the creation of an accurate model, another approach and the one that this study focuses on is from the pretrained model itself, a finetuner or additional classifier in place for the purpose of specific facial recognition. D. Li et al. published a paper on multiple-step model training method for Asian faces [12]. With the model achieving high accuracy in real world application, this approach towards solving the problem of face recognition has proven to be promising. Despite differences in methodology, our approach to the solution is the same as we strive to improve the existing models.

## III. MODEL FOR FACE RECOGNITION

This project employs use of Histograms of Oriented Gradients (HOGs) with a trained Support Vector Machine (SVM) as basis for a face detector [13]. To improve the accuracy of the recognition model, we add a face alignment function to the detection model. The alignment is largely based on the facial landmarks model that has been previously trained with manual annotation on images of faces through the amazing efforts of the i-bug group [14]–[16]. For this project, we used 68 facial landmarks for the alignment model as it performed the best within the spectrum of our computational limitations. There are models of up to 194 landmarks that produce even faster and more accurate results [17]. The face recognition model is based on the DCNN model of FaceNet [3] to create 128 embeddings for each images. However, instead of using a loss function to calculate the distance between faces to verify if they are the same person, we use a trained kNN classifier for recognition and using the embeddings as inputs. Figure 2 represents the architecture model of this project.

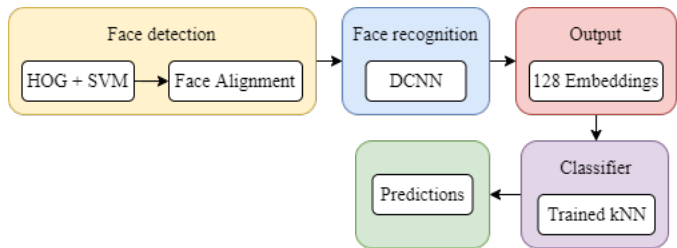


Figure 2. Face recognition architecture model of receptionist robot

### A. Face Detection

Face detection became popular and reached confident heights when the Viola-Jones object detector came about [18] by applying Haar-like features alongside Adaboost to make up a rapid cascade classifier. From then on, there have been numerous methods created for facial detection in both machine learning and deep learning. Most prominent in machine learning



for face detection is HOGs and in deep learning is a multi-task convoluted neural network (MTCNN) [19]. For this project, we have chosen to use HOGs as a face detector due to the fact that this algorithm has been extremely stable and reduces any workload that might be bottleneck when applying to a different system.

In spite of the fact that HOGs were created in the mid-2000s, its capabilities for facial detection are still relevant even in today's standards. Figures 3 and 4 show how facial landmarking is done where many faces are present and for a single face respectively. However, such technology is not without its drawback. When a face is angled at a certain degree from center facing, the alignment algorithm has a hard time making the detected face fits the landmarks and this in turn affects the accuracy of the recognition process. One of the harder angles for the detector to align is shown in Figure 5.



Figure 3. Face detector running on multiple faces

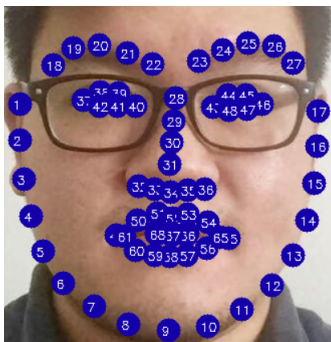


Figure 4. 68 landmarks annotated on a single face

### B. Face Recognition

The recognition process takes place after the faces have been detected by a DCNN model. For this particular project, we are employing the pre-trained recognition model of FaceNet as a tool for creating image embeddings. This system, built in 2015, boasted an accuracy of 99.63% on the LFW Challenge by creating 128 embeddings for each image and using a loss function to calculate the Euclidean distance between pairs of images to decide whether they are similar or not. As mentioned before, the loss function provided extremely high accuracy for

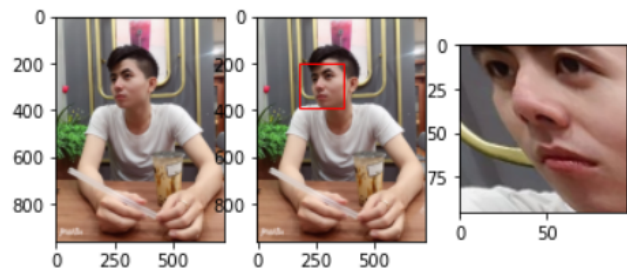


Figure 5. Face alignment of a person looking at a difficult angle

LFW due to the usage of the LFW training dataset as well as the Youtube Faces dataset, and despite having a standard testing threshold of 0.6, could not differentiate Asian faces where the distances can span towards the low end of 0.2. Figure 6 describes the model structure of the FaceNet system.

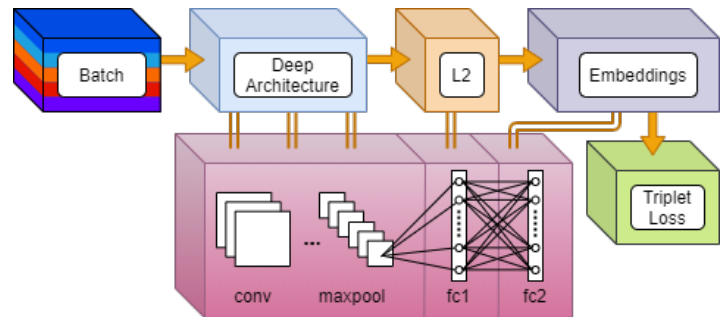


Figure 6. Model structure of FaceNet architecture [3]

The formation of embeddings that are used for recognition is done through a process of creating convolutional layers and connecting them together to make a feature map for an image. This type of DCNN is called a Siamese neural network or twin neural network, and it is heavily applied in One-shot learning technologies [20]. At the beginning of the training process for the neural network, an image is passed through multiple kernel filter that focuses on identifying a specific feature of the image. The resulting images that have passed through are now convoluted layers and are subjected to max pooling, an action that calculates the maximum value of these feature maps then pools them together. Max pooling is meant to downsize the resulting layers and highlight the features. This process is done multiple times until only a pool the highlighted features remain of the original image.

These resulting layers are then interconnected to create the first fully connected layer, this is known as flattening. Since the feature layers are matrices, flattening makes them into a single vector input for the next step. However, before that, the model has only gone through the forward pass only once and this is not enough to create a sufficient model. Back propagation is done to re-adjust the weights and biases of the model. The act of completing the forward pass and back propagation is an epoch. After multiple epochs are cycled through and the learning curve converges, the training process is considered



finished. The FaceNet [3] model learning curve converges after 500 hours of training and had a total of 140 million parameters between 22 layers. The last layer is where embeddings are calculated as distances, this is the L2 normalization layer or Euclidean normalization layer. The differences between the two images are then calculated via a loss function. If the distance between the embeddings of an image and the embeddings of the anchor is below the set threshold then they are the same, relatively speaking.

With this project, as previously mentioned, we will not be employing Euclidean normalization to compare and verify the similarities between the images but rather to verify whether the faces viewed by the system are amongst the previously trained faces or not. It is also important to understand that the differences between faces with and without Asian features are not as apparent as they might seem. Due to the way the features of the face are created by the algorithm, we currently cannot grasp the logic behind which features are chosen and why. Thus, making it extremely difficult to know figure out the reason why Asian faces are harder to differentiate. However, it would only be a problem if we were to run the algorithm independently. By adding a classifier at the end of the process of features creation as the recognizer, we can work around the problem without much changes to the original algorithm. And the de facto recognition process is done by the kNN classifier.

### C. K Nearest Neighbor Classifier

K Nearest Neighbor Classifier has long become a staple in the arsenal of the machine learning community with its ease of implementation and debugging. One of the benefits that kNN brings to this project as we apply it in place of the loss function is the capability to classify multiple dimensions [21]. Aside from the ease of usage, all the kNN does with the embeddings is with the trained data, neighbors will take majority votes in the feature maps and assign a class to the embeddings. And so, this accomplishes our goal of recognizing Asian faces with relatively light workloads. With the loss function used as a known and unknown filter, it reduces the likelihood of false-positive happening when verifying the faces with the classes in the database. Figures 7, 8, 9 is the 2D visual representation of the kNN at three different perplexities. Since the data is consisted of 128 dimensions, these figures serve as visualizations for the multi-dimensionality that the kNN classifier is working on.

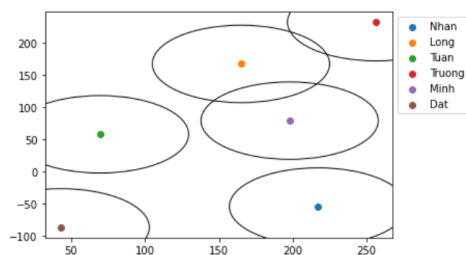


Figure 7. K Nearest Neighbor 2D visualization at perplexity 1

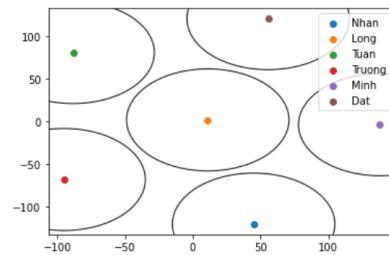


Figure 8. K Nearest Neighbor 2D visualization at perplexity 2

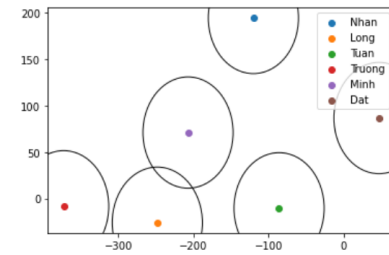


Figure 9. K Nearest Neighbor 2D visualization at perplexity 3

One of the advantages that kNN has over other classic types of classifiers is that the characteristic of kNN allows for an accurate recognition rate between known faces and natural unknown faces as opposed to SVM. While SVM can make for a precise recognition system, it cannot recognize unknown faces. When the data is captured through the camera and inputted into the system, SVM will only separate the data and assign it between the known classes even if they do not belong in any of them. To combat this, we need to assign a separate "Unknown" class and have a massive amount of data exponentially larger than our known data to have an accurate prediction. And thus, by going through all of these processes, the system will need more time and computational power for each face. Despite eventually yielding more better results, the cost efficiency of this method is not practical for a real-world product. A more advanced version of SVM, ArcFace [2], was made so that the flaws of having to create a separate unknown class and training a massive amount of data are irrelevant.

### IV. STANDARDIZED METHOD OF DATA COLLECTING

Data collecting is one of the most important defining factors in any deep and machine learning endeavors. There are two ways of data collection, mass data gathering and data augmentation. Mass data gathering is essential for creating a better face recognition model and the more data is being learned the more accurate the model will be. However, this takes tremendous efforts and very high computational capabilities so with our small amount of data, we can augment with various tool to create many variations of the same image. When taking our project into account, there are major hurdles that need to be addressed for any chance of real world implementation. One of the key factors that affects the recognition process and reflect heavily on its accuracy is the various environmental lighting.

As a receptionist and security robot, controlled lighting settings that are well lit and noise free are not always the case. And if a robot fails under non optimal conditions then it is not real-world ready. For this very purpose, we have create our dataset with three illumination degree: brightly lit, natural warm light and half shaded. Figure 10 shows the face being capture in half shaded poor lighting condition. A study done on data collection for facial recognition [22] also pointed out an outside factor that greatly affects recognition accuracy is the heights variation of the observee.

Since the levitation angle of the camera for the receptionist robot is fixed, even if the data is of clear and well define frontal faces, the angle of observation can make quite a difference when processing the inputting image. A face viewed from certain angles can make the identity perception for it completely different, that is why data from any individual face have to be taken at different angles. Following the study by F. R. Maciel et al. [22], the angle of the face tilted from the frontal view is from 15° to 30°. This ensure that most of the angles of capture that the face recognition system can be optimized for do not exceed the angles of detection and can function properly when faced with high variations between individual heights. Shown in Figure 11 are examples of the images that were made via this method and used in the testing dataset.



Figure 10. Images of a face captured in half shaded condition



Figure 11. Images of a face captured in different angles

With the methodology being done, there was still a problem in terms of real world application, the time needed to collect data from a single person when adding new inputs individually or in batches. When applied to a working environment, a receptionist and security robot need to be concise and time-efficient when adding new data. And since the action of gathering data from a persons face is currently being done manually, we are applying image sequencing in order to create a mass of data in a short period of time. The data collection is done in an environment where lighting can be changed to mimic different conditions and a video of roughly 15 seconds will be taken for each condition per person. At an average time taken of 2 to 3 minutes, a video recording of 60fps, and taking

every 10 frames, we get a sum of 270 images per person. This amount of data for each face ensures that in the duration of the robot operation, the margin for error is minimal. Figure 12 is the visualization of the trained data of the kNN classifier.

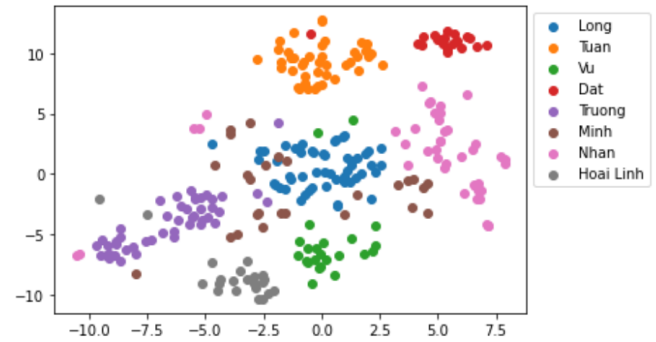


Figure 12. 2D visualization of all the trained data

## V. RESULTS

When all elements of the project are in place, the test is done with a dataset which consists of 1020 individual Vietnamese celebrities with 10 known faces from our own set. The number of images netted to 3500 in total. Figures 13 and 14 shown are the distances between negative and positive pairs when the dataset is tested by our system respectively. The total negative pairs are roughly 150M and the positive pairs are around 200K. Right now, the threshold is chosen as 0.45 as an approximation. However, with the data that we have worked with, it became even more apparent that the differences between faces are extremely subtle as represented by the smooth transition in the graph along the distance axis. As we become more consistent with our data gathering and classification, a density distribution function is needed to optimize the threshold for maximizing accuracy.

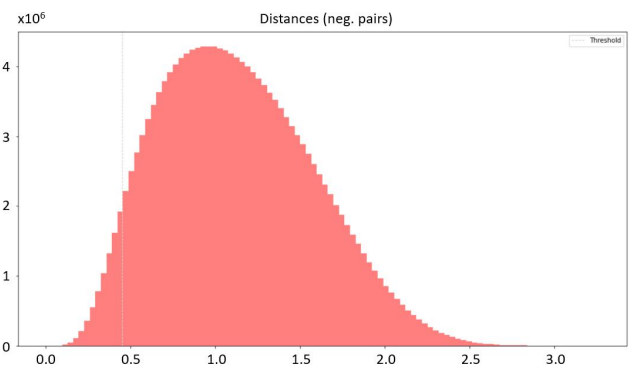


Figure 13. Histogram of negative pairs during VN celebrities test

The accuracy of this system is shown in Table I. Accuracy is calculated based on the number of frames that a face is detected and the number of frames that the right name for that face is counted. The test is done for 3 separate datasets: a single image passed through the loss function as a One-shot learning

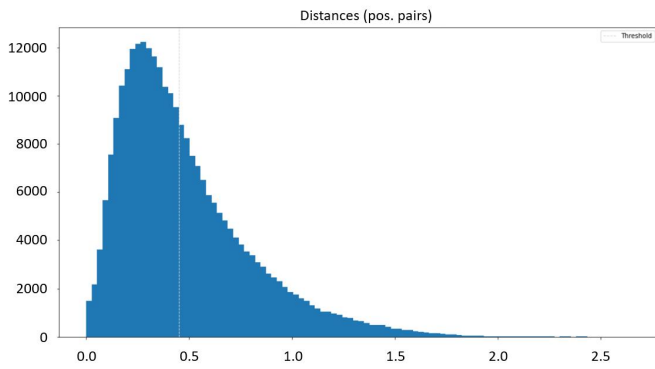


Figure 14. Histogram of positive pairs during VN celebrities test

model; a dataset where each class has a randomized collection of images; a dataset that applies the data collecting method. The shown results are averages calculated for 10 known classes in the project. The test is done for the purpose of real-time application and during many intervals of time of the day where lighting condition differs

Table I

AVE. ACC. OF DIFFERENT DATASET MODEL IN DIFFERENT LIGHTING CONDITIONS

Dataset Models	Ave. correct frames	Ave. acc. (%)
One image dataset (w/o classifier)	80357/120000	66.96
Randomized datasets	96521/120000	80.43
Methodized datasets	104423/120000	87.01

The accuracy of kNN when performing classification is shown in Table II. Since the unknown classification is done before the actual face recognition process, the accuracy is significantly lower than when done with different images of the learned classes.

Table II

AVERAGE PERFORMANCE ACCURACY OF DIFFERENT TEST MODELS IN REAL TIME

Test models	Ave. correct frames	Ave. acc. (%)
Standalone face detection	119986/120000	99.98
Verification between known classes	112355/120000	93.62
Verification with unknowns	104423/120000	87.01

Lastly, with the system running smoothly, we now apply it to our robot model. The creative choices of the design derive from the simplistic design of HAL9000 and as our way to pay homage to the film that inspired us. The real-time testing on the actual model is done on a Jetson TX2, which was the fastest standalone CPU that was available to us. The entire system runs at 5 fps on average with the Jetson TX2 and at 16 fps on average with an NVIDIA GTX 1050 on our laptops. Despite the low frame rates, the model still yields extremely positive results as mentioned above and is susceptible to even more upgrades to the design and hardware should need be. Figure 15 shows the Inventor design and the built model of the case, and Figure 16 is the system running with live feed videos.



Figure 15. Design of the actual robot model

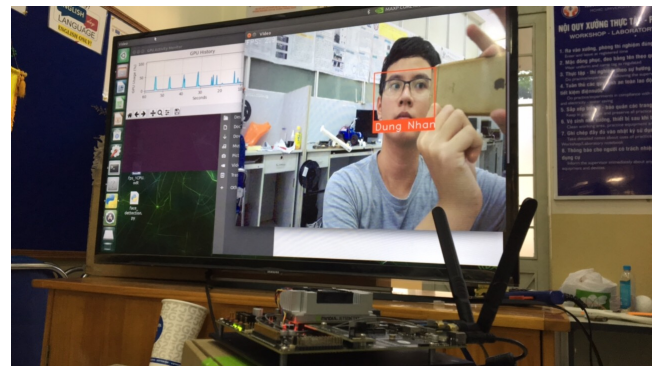


Figure 16. Live video feed testing

## VI. CONCLUSION AND FUTURE PROSPECT

As a solution for recognition between Asian faces, this system has provided sufficient accuracy regarding determining unknown faces and differentiate between the known faces. With the goal of real-world application in mind, the result is a significant improvement from previous models and prove the feasibility of face recognition based receptionist and security robot. This not only serves as an example of the capabilities of face recognition technology but also a small milestone in solving the problem of facial feature variations from people of different races. But for now, our project has successfully addressed the problems of similarity between Asian faces.

In spite of the shortcomings that this project has, the potential for further development is vast. One of the most important updates that we will make to the system in the future is creating its own unique model based on the massive dataset that Z. Xiong et al. [11] have made. Implementing this face recognition model to the arsenal of the ever-growing human-robot interactive technology and making machines more human-like. Furthermore, to develop an algorithm that accurately selects the threshold the current goal to strive for at this time. Last and certainly not least, score some small victories for humanity in the efforts to push the limits of technological advancements.

## ACKNOWLEDGMENT

We would like to thank the FaceRec and CompVis communities for all the ideas and help that were given our way. To have done so much in so little time was quite overwhelming, so thank you, kind strangers. Also much love to the class of 16142CLA, our classmates, and Vu-Thien Le for their supports in creating the data for this project and for tolerating our constant demands for more data.

## REFERENCES

- [1] M. Coşkun, A. Uçar, Ö. Yildirim, and Y. Demir, “Face recognition based on convolutional neural network”, in *2017 International Conference on Modern Electrical and Energy Systems (MEES)*, 2017, pp. 376–379.
- [2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, *Arcface: Additive angular margin loss for deep face recognition*, 2018. arXiv: 1801.07698 [cs.CV].
- [3] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering”, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015. DOI: 10.1109/cvpr.2015.7298682. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2015.7298682>.
- [4] A. J. Shepley, *Deep learning for face recognition: A critical analysis*, 2019. arXiv: 1907.12739 [cs.CV].
- [5] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kennsuke, “Cultural differences in attitudes towards robots”, Jan. 2005. DOI: 10.13140/RG.2.2.22507.34085.
- [6] S. Šabanović, “Inventing japan’s ‘robotics culture’: The repeated assembly of science, technology, and culture in social robotics”, *Social Studies of Science*, vol. 44, no. 3, pp. 342–367, 2014, PMID: 25051586. DOI: 10.1177/0306312713509704.
- [7] A. K. Pandey and R. Gelin, “A mass-produced sociable humanoid robot: Pepper: The first machine of its kind”, *IEEE Robotics & Automation Magazine*, vol. PP, pp. 1–1, Jul. 2018. DOI: 10.1109/MRA.2018.2833157.
- [8] P. Holthaus and S. Wachsmuth, “The receptionist robot”, Mar. 2014. DOI: 10.1145/2559636.2559784.
- [9] F. Bazzano and F. Lamberti, “Human-robot interfaces for interactive receptionist systems and wayfinding applications”, *Robotics*, vol. 7, p. 56, Sep. 2018. DOI: 10.3390/robotics7030056.
- [10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments”, University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.
- [11] Z. Xiong, Z. Wang, C. Du, R. Zhu, J. Xiao, and T. Lu, “An asian face dataset and how race influences face recognition”, in *Advances in Multimedia Information Processing – PCM 2018*, R. Hong, W. Cheng, T. Yamasaki, M. Wang, and C. Ngo, Eds., Cham: Springer International Publishing, 2018, pp. 372–383, ISBN: 978-3-030-00767-6.
- [12] D. Li, X. Zhang, L. Song, and Y. Zhao, “Multiple-step model training for face recognition”, in *International Conference on Applications and Techniques in Cyber Security and Intelligence*, J. Abawajy, K. R. Choo, and R. Islam, Eds., Cham: Springer International Publishing, 2018, pp. 146–153, ISBN: 978-3-319-67071-3.
- [13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, 2005, 886–893 vol. 1.
- [14] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge”, in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.
- [15] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: Database and results”, *Image and Vision Computing*, vol. 47, Jan. 2016. DOI: 10.1016/j.imavis.2016.01.002.
- [16] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “A semi-automatic methodology for facial landmark annotation”, in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 896–903.
- [17] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees”, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1867–1874. DOI: 10.1109/CVPR.2014.241.
- [18] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features”, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, Dec. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.
- [19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks”, *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, ISSN: 1558-2361. DOI: 10.1109/lsp.2016.2603342. [Online]. Available: <http://dx.doi.org/10.1109/LSP.2016.2603342>.
- [20] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition”, in *ICML deep learning workshop*, Lille, vol. 2, 2015.
- [21] P. Cunningham and S. Delany, “K-nearest neighbour classifiers”, *Mult Classif Syst*, Apr. 2007.
- [22] F. Maciel, S. Cleger-Tamayo, A. Khurshid, and P. Martins, “Data collection and image processing tool for face recognition”, in Sep. 2019, pp. 386–392, ISBN: 978-3-030-30711-0. DOI: 10.1007/978-3-030-30712-7\_49.



# An efficient evolutionary algorithm for joint optimization of maintenance grouping and routing

Ho Si Hung Nguyen

*Faculty of electrical engineering*

*The University of Danang - University of Science and Technology*

Danang, Vietnam

nhshung@dut.udn.vn

Phuc Do

*CRAN, UMR 7039, Campus Sciences, BP 70239*

*University of Lorraine*

Nancy, France

phuc.do@univ-lorraine.fr

Hai-Canh Vu

*Mechanical Engineering Department*

*The University of Technology of Compiègne*

Compiègne, France

hai-canh.vu@utc.fr

Thanh Bac Le

*Faculty of electrical engineering*

*The University of Danang - University of Science and Technology*

Danang, Vietnam

lethanhbac@ac.udn.vn

Kim Anh Nguyen

*Faculty of electrical engineering*

*The University of Danang - University of Science and Technology*

Danang, Vietnam

nkanh@dut.udn.vn

**Abstract**—Geographically Dispersed Production System (GDPS) is a system network in which production sites are located far away from each other and subjected to management of an administration center. In maintenance planning for such system, planners intend to group maintenance activities to save preparation cost and transportation cost. Specially, this work is extremely meaningful in framework of sustainability (reduction of environmental pollution). In fact, when grouping maintenance activities of components located at different sites, maintenance resources need to be travelled from the maintenance center to the production sites where these components are located on, and then come back to the maintenance center when the maintenance is completed. By that way, travel distance as well as emissions from means of transports can reduce compared to individual maintenance. Although grouping maintenance has the positive impacts, it causes difficulties due to the shift of maintenance dates of components. Moreover, maintenance planning optimization for a GDPS is multi-groups and multi-routes problem. Hence, optimizing maintenance grouping and routing for GDPS should be considered simultaneously. However, the joint optimization leads to be a NP-hard problem to be solved. To face this issue, in this paper an evolutionary algorithm based on Genetic Algorithm (GA) and Branch and Bound algorithm (BAB), named GA-BAB, is developed. The numerical study shows that the proposed algorithm can efficiently solve maintenance grouping and routing optimization problem in a GDPS context.

**Index Terms**—Geographically dispersed production system, genetic algorithm, branch and bound algorithm, efficient evolutionary algorithm.

## I. INTRODUCTION

In the recent years, to improve the competitiveness, many manufacturing companies have changed their production mod-

els to adapt to the market. In particular, to adapt to geographical dispersion of customers, a manufacturing system, named geographically dispersed production system (GDPS), is emerged. The GDPS is a distributed manufacturing system in which production sites are located far apart from each others. The production sites have to be connected and subjected to management of an administration center. The production sites are subsystems constructed by multiple components. In real applications, the GDPSs have to face many challenges concerning standards, regulation, production management, and especially, maintenance planning and optimization due to the geographical dispersion of production sites [12].

In fact, planning a maintenance strategy for a GDPS is a very complex work due to the dispersed location of production sites impacting on optimization of maintenance routing [1]. In addition, maintenance for a GDPS is impacted by some specific constraints such as disruption, possibility of accident, weather or road condition, etc [5]. Recently, some maintenance strategies have been developed for GDPS systems, see [1], [5]. However, these works mainly focus on the problem of maintenance routing rather than maintenance planning and optimization ones. Recently, paper [10] removes the assumption about the number of PM activities and builds two different models for maintenance planning and maintenance routing. The maintenance planning model based on grouping maintenance strategy considers joint maintenance of several components to save preparatory cost. While, the maintenance routing model considers joint maintenance of components in an adequate maintenance itinerary to reduce travel distance, travel time



and transportation cost compared to individual maintenance. However, the existing work has some following common limitations: (a) maintenance route and grouping maintenance plan are not rescheduled based on the new situation of the components/GDPS; (b) optimization of maintenance routing and maintenance grouping are sequentially implemented. From our best knowledge, joint optimization of maintenance grouping and routing has not yet considered for a GDPS. Thus, an optimization algorithm for joint optimization of maintenance grouping and routing has not yet developed also. This open issue motivates our works presented in this paper.

In relation to optimization of maintenance grouping, many optimization algorithms have been developed to a grouping maintenance plan with minimum maintenance cost or maximum economic profit. When components are fixed in groups, many studies address optimal maintenance planning as an optimization problem of maintenance routing by exact algorithms [8]. From optimization point of view, exact methods are appropriate to solve maintenance routing problems. It is completely true in cases that the number of production sites is not too large, where computation time by exact methods is not too longer than heuristic methods. The exact methods are mentioned in literature as Exhaustive Search [8], Branch and Bound [9], Branch and Cut [4], etc. In these methods, Branch and Bound (BAB) method emerges as effective methods due to (a) reducing computation time compared to exhaustive search thanks to trying to rule out parts of the search space that cannot contain the best solution; (a) easily to simulate [4].

The exact algorithm always ensures that obtained results are global optimal ones. However, the computation time of this method grows up exponentially if the number of components, sites, maintenance activities or constraints increases. Specially, when the components can randomly combine to create groups, the number of group combinations is extremely larger. Therefore, using the exact algorithm is impossible due to the too long computation time. To deal with this NP-hard problem, a natural solution is to use a genetic algorithm (GA) [6]. It has been used and considered as a relevant optimization approach in the context of maintenance optimization [2], [7], [11], [13]. The GA, a heuristic algorithm, proves effectiveness in solving maintenance planning optimization problem for large-scale system including multi-components and multi-constraints. The aim of the algorithm is to find the best solution in reasonable computation time instead of the global optimal one. Besides, speed and duration of computation time can be adjusted by changing iteration or the special parameters of GA. Of course, the accuracy level of obtained results is proportional to computation time. That is the strong point of this method compared to analytical and exact methods. This strength can be raised many times if GA can be combined with other algorithms. From this point of view, we develop a GA-BAB algorithm by using GA and Branch and Bound algorithm with objective to maximize grouping economic profit.

The remainder of this paper is organized as follows. In Section 2 we describe the problem and establish a cost model. It is a base to build a grouping economic profit mode. After

that, in Section 3 we present how to find a maintenance itinerary of a specific group based on Branch and Bound algorithm and then, a GA-BAB using GA and Branch and Bound algorithm is developed to optimize grouping maintenance planning. Section 4 provides some computational results when deploying the proposed algorithm with objective to maximize grouping economic profit. Finally, in Section 5, we present our conclusions and discuss future research.

## II. PROBLEM DESCRIPTION AND MATHEMATICAL FORMULATION

### A. Definition of the problem and working assumptions

We consider a GDPS with  $n$  components located at  $m$  sites. The production of each site is supported by several machines (components) subjected to random failures. The failure rate of these components is assumed to be increasing over time by Weibull distribution law. To prevent the system from an intensive regime of failures, preventive maintenance (PM) is carried out at predetermined times. A component is considered to be in “as good as new” state after the PM. Whereas, corrective maintenance (CM) is immediately carried out after the component failures in order to restore the failed components into their operational state as soon as possible. After a minimal repair action, component is in “as bad as old” state. It should be noted that in this paper, we consider that to perform a CM action, local resources (repair team, repair tool and spare part) at maintenance sites are enough. Otherwise, to do a PM action, the external maintenance resources from the maintenance center are needed. For this reason, PM is performed jointly on a group of components situated in different sites. To do that, the maintenance resource has to travel from the maintenance center to the maintenance sites where the components are located on, and then come back to the maintenance center when the maintenance of all components in the group is completed. By that way, this average total cost includes the cost of maintenance actions (PM and repairs), the transportation cost, and penalty costs corresponding to late arrivals or soon (compared to tentative plan). The objective is to determine simultaneously the optimal routing sequence and the optimal PM schedule in order to minimize the expected total cost/ maximize economic profit, considering distance constraint. The following assumptions are considered:

- A maintenance itinerary starts and ends at a global maintenance center.
- Failure of a component leads the production site to stop.
- Interruption of a production site does not impact the normal operation of the others.
- Only one maintenance team (repairman) is considered.
- The maintenance team and resources are always available and ready for the maintenance.
- For each component, different levels of technician's skill are required
- CM duration is neglected when compared to the length of the planning horizon

Assume that a group of several components, denote  $G_k$ , are preventive maintenance together. These components of group  $G_k$  located at  $m_s$  sites are tentatively maintained at their individual PM dates  $t_{im}$ . To be jointly maintained, the individual PM dates have to be adjusted, and the components of  $G_k$  are actually preventively maintained at  $t'_{im}$  instead of  $t_{im}$ , with  $i \in G_k$ . Let  $t_{G_k}$  denote the time that the maintenance team leaves the maintenance center (departure time) to maintain group  $G_k$ , and  $I_{G_k}(j) = v$  means that with respect to the route  $I_{G_k}$ , site  $j$  is visited by the maintenance team at the  $v^{\text{th}}$  order

- If  $I_{G_k}(j) = 1$ , the maintenance team will visit site  $j$  at first after leaving the maintenance center, the maintenance team will arrive at site  $j$  at  $TAS_j = t_{G_k} + t_{0j}^{tr}$ , where  $t_{0j}^{tr}$  is the travel time from the maintenance center to site  $j$ .  $TAS_j$  is the grouped PM date of components of group  $G_k$  located at site  $j$ :  $t'_{im} = TAS_j$ . The maintenance team will leave site  $j$  at time  $TLS_j = TAS_j + \sum_{i \in G_k; i \in j} \omega_i$ , and move to the next maintenance site.
- If  $I_{G_k}(j) = v > 1$ , the maintenance team will visit site  $j$  at the  $v^{\text{th}}$  order. The actual PM dates of components of group  $G_k$  located on site  $j$  are then  $t'_{im} = TLS_q + t_{qj}^{tr}$ , where  $q$  is the site that the maintenance team has previously visited to at the  $(v - 1)^{\text{th}}$  order.

#### B. Cost model and grouping economic profit mode

The objective of this section is to formulate a cost model of components performed jointly preventive maintenance. Based on this model, grouping economic profit model is formulated by comparing to individual maintenance cost.

1) *Cost model*: The maintenance cost of group  $G_k$  contains the following parts: spare part cost ( $C_{G_k}^{sp}$ ), downtime cost ( $C_{G_k}^{dt}$ ), labor cost ( $C_{G_k}^{lb}$ ), site preparation cost ( $S_{G_k}^0$ ), and travel cost ( $S_{G_k}^{tr}$ ).

$$\begin{aligned} C_{G_k}^p &= C_{G_k}^{sp} + C_{G_k}^{dt} + C_{G_k}^{lb} + S_{G_k}^0 + S_{G_k}^{tr} \\ &= \sum_{i \in G_k} C_i^{sp} + \sum_{i \in G_k} R_i^{dt} \omega_i + R_{G_k}^{lb} (l_{\max}) \cdot \sum_{i \in G_k} \omega_i \\ &\quad + \sum_{j \in G_k} S_{ji}^0 + R^{tr} \cdot L_{G_k}(I_{G_k}) \end{aligned} \quad (1)$$

where,  $\omega_i$ ,  $C_i^{sp}$ ,  $C_i^{dt}$  and  $C_i^{lb}$  are preventive maintenance duration, spare part cost, downtime cost and labor cost for a PM activity of component  $i$ ;  $S_{ji}^0$  is site-preparation cost of a component  $i$  at site  $j$ ;  $S_{ij}^{tr}$  is transportation cost from a global maintenance center to site  $j$  containing maintained component;  $R^{tr}$  is transportation cost rate;  $l_i$  is skill level of a maintenance team required by component  $i$ ;  $R_i^{lb}(l_i)$  is labor cost rate of maintenance team corresponding to skill level  $l_i$  and  $R_{G_k}^{lb}(l_{\max})$  depends on  $l_{\max} = \max_{i \in G_k} l_i$ ;  $L_{G_k}(I_{G_k})$  is total travel distance determined by itinerary maintenance  $I_{G_k}$ .

2) *Grouping economic profit model*: To evaluate the effectiveness of grouping maintenance strategy, grouping economic profit is normally used. The grouping economic profit is defined as the difference between the total maintenance costs that have to be paid when the components are grouped and

when they are separately maintained. The grouping economic profit of group  $G_k$ , denoted  $EPG$ , can be expressed as follows:

$$\begin{aligned} EPG_{G_k}(t_{G_k}, I_{G_k}) &= \sum_{i \in G_k} C_i^p - C_{G_k}^p - \Delta H_{G_k} \\ &= \left( \sum_{i \in G_k} S_i^0 - S_{G_k}^0 \right) + \left( \sum_{i \in G_k} S_i^{tr} - S_{G_k}^{tr} \right) \\ &\quad - \left( C_{G_k}^{lb} - \sum_{i \in G_k} R_i^{lb} \cdot \omega_i \right) - \Delta H_{G_k} \\ &= \Delta S_{G_k}^0 + \Delta S_{G_k}^{tr} - \Delta C_{G_k}^{lb} - \Delta H_{G_k} \end{aligned} \quad (2)$$

where,

- $\Delta S_{G_k}^0$  is the site-preparation cost saving when several components of the same site are preventively replaced  $S_{G_k}^0 = \sum_{j \in G_k} (nc_j - 1) S_j^0$ ;  $nc_{jk}$  denote the number of components of group  $G_k$  located at site  $j$ ; site-preparation cost is only paid once time if many components at the same site are maintained jointly.
- $\Delta S_{G_k}^{tr} = \sum_{i \in G_k} S_i^{tr} - S_{G_k}^{tr}$  is the travel cost saving.  $\sum_{i \in G_k} S_i^{tr}$  is the total travel cost when the components are maintained separately. It means that the maintenance team travels from the maintenance center, then performs the PM on only one component at a time, and returns to the maintenance center. This cost is expressed as follows.

$$\sum_{i \in G_k} S_i^{tr} = R^{tr} \cdot \sum_{j \in G_k} \sum_{i \in j} (D_{0j} + D_{j0})$$

- Labor penalty cost is the difference between the labor cost paid for the maintenance of group  $G_k$  and that of its components. We consider that the labor cost rate of components and a group of components may be different.  $\Delta C_{G_k}^{lb} = R_{G_k}^{lb}(l_{\max}) \cdot \omega_{G_k} - \sum_{i \in G_k} R_i^{lb}(l_i) \cdot \omega_i$ .
- $\Delta H_{G_k}$  is the penalty cost occurred due to the change of the maintenance dates of the components in the group. Note that to be grouped in the same group, the maintenance dates of individual components in the group have to be modified. The detail calculations of this penalty cost can be found in [3].

In a short-term planning horizon, there are many PM activities and a grouping solution of this horizon therefore may contain several groups. A grouping solution or grouping structure can be defined as a collection of mutually exclusive groups  $SG = \{G^1, G^2, \dots, G^e\}$  with  $G^h \cap G_k = \emptyset$  and  $G^1 \cup G^2 \cup \dots \cup G^e$  covers all maintenance activities in the planning horizon. The total economic profit of grouping structure  $GS$  can be calculated as .

$$EPS(GS) = \sum_{G_k \in GS} EPG_{G_k}(t_{G_k}, I_{G_k}) \quad (3)$$

$EPS(GS)$  represents the performance of grouping structure  $GS$ .

### III. EVOLUTIONARY OPTIMIZATION ALGORITHM

The aim of this section is to develop an efficient evolutionary algorithm to optimize maintenance planning for a GDPS. The maintenance plan ensures that travel distance is minimum while economic profit is maximum. To do that, the optimization process can be divided into two following phases: (i) optimization at group level of components to maximize the economic profit of a group (*EPG*) with distance constraint (minimum travel distance); (ii) optimization at grouping structure level where all groups of components of a grouping solution are considered to maximize the economic profit the considered short-term horizon (*EPS*). For this purpose, at group level, a modified Branch and Bound algorithm (*M-BAB*) is developed to find the shortest maintenance itinerary of group with maximum *EPG*. In while, at the grouping structure level, an efficient evolutionary algorithm based on Genetic Algorithm (GA) and M-BAB algorithm, named GA-BAB, is developed to find the best grouping structure with satisfying (i) maintenance itineraries of groups is the best with minimum travel distance; (ii) grouping economic profit (*EPS*) is maximum.

#### A. Modified Branch and Bound algorithm

In this subsection, the objective is to develop M-BAB algorithm. When travel cost rate ( $R^{tr}$ ) is constant, the shortest maintenance itinerary means that transportation/travel cost is the lowest. We assume that a group is known. We find the optimal itinerary for the maintenance team to do the maintenance of all components of the group at the lowest penalty and transportation costs. To do this, both maintenance routing and maintenance scheduling optimization are considered in search process of the M-BAB algorithm. Firstly, the algorithm has to find an optimal maintenance itinerary satisfying the the lowest transportation cost. After that, the maintenance scheduling is optimized based on this itinerary.

To facilitate the understanding, let us go to an example of a search tree using M-BAB algorithm as in Figure 1. In the Figure, the root node (node 0) at the top of the tree represents the global maintenance center, the nodes at level one represent all the sites that could be visited first (node 1, node 2, and node 3), the nodes at level two represent all the sites that could be visited second (node 4, node 5, node 6, and node 7), etc. Generally, the horizontal set of all nodes at level  $l$  is denoted by  $HN_l$ . Otherwise, the vertical set of nodes, denoted  $VN_{a \rightarrow b}$ , represents an entire route ( $a \equiv b \equiv 0$ ) or a part of route from node  $a$  to node  $b$ . For example, in Figure 1,  $VN_{0 \rightarrow 8}$  denotes the part of route where the maintenance team travels from the global maintenance center to node 3, node 6, and node 8 consecutively. The M-BAB algorithm iteratively solves the maintenance routing problem by considering first level at a time starting from the top of the search tree (level 0) to the last level (level  $ns_k + 1$ ). In Figure 1, we consider the maintenance routing of group  $G_k$  containing components 1, 2, 4, 5, 7, and 8 located in three different sites 1, 2 and 3. The BAB will be done from the maintenance center (level 0) to the last level (level 4).

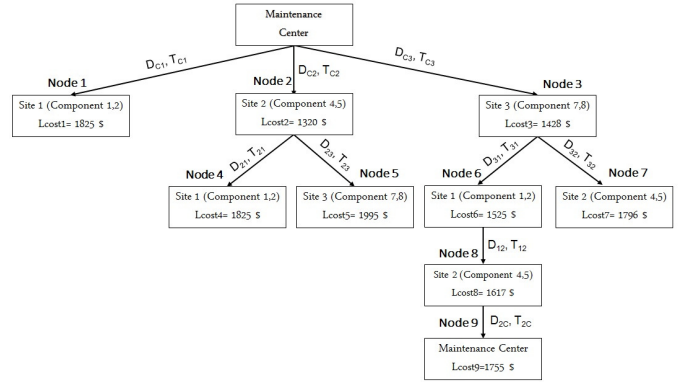


Fig. 1. Search tree for the group  $G_k = \{1, 2, 4, 5, 7, 8\}$

At a considered level  $l$ , the following steps are realized:

- 1) Estimate lower bound value of each node  $q \in HN_l$ . The lower bound value of node  $q$ , denoted  $LB(q)$ , indicates the minimum logistic cost (minimum travel distance), that we can obtain if we decide to travel across the node  $q$ .
- 2) Select a node for expansion. Let  $EHN_l$  be the set containing all nodes of level  $v$  ( $HN_v$ ) and all unexpanded nodes of the previous levels. The node with the lowest value of the lower bound among all nodes in  $EHN_l$  (the most promising node) will be selected for the expansion.
- 3) Expand the selected node. The node expansion is to identify all possible nodes that the maintenance team can directly travel to from the selected node. The expanded node is called parent node, and the generated nodes are child ones. The BAB process then jumps to the level of the child nodes.

The only one problem now is how to estimate the lower bound value. Firstly, let  $D_{jm}$  denote the distance between two sites  $j$  and  $m$ .  $D_{jm} = +\infty$  if there is no direct route connecting the two sites. The lower bound of transportation cost which represents the smallest transportation cost, that the routes traveling across node  $q$  can be obtained, is calculated as

$$LBTC(q) = R^{tr} \cdot (D_{0 \rightarrow q} + \hat{D}_{q \rightarrow 0}) \quad (4)$$

- $D_{0 \rightarrow q}$  is the total travel distance of the planned part of the routes from the maintenance center to node  $q$ . Consider node 2 in Figure 1, we have  $D_{0 \rightarrow 2} = D_{02}$ .
- $\hat{D}_{q \rightarrow 0}$  is the expected total travel distance of the unplanned part of the routes from node  $q$  to the maintenance center. Since the route from node  $q$  to global maintenance center is still unknown, the calculation of its total distance has to be done approximately. Consider a site  $m$  that the maintenance team has not yet visited  $m \notin VN_{0 \rightarrow q}$ , it is clear that the shortest distance to go to site  $m$  is  $D_m^{min} = \min_{j \in S, J} D_{jm}$ .  $SJ$  is the set of sites that the maintenance team could be in before visiting site  $m$ .

Consequently, we have

$$\hat{D}_{q \rightarrow 0} = \sum_{m \notin VN_{0 \rightarrow q}} D_m^{min} \quad (5)$$

For example, turning back node 2 as shown in Figure 1, the maintenance team has already visited site 2. The expected travel distance of the unplanned part is

$$\hat{D}_{2 \rightarrow 0} = \min(D_{10}, D_{30}) + \min(D_{21}, D_{31}) + \min(D_{23}, D_{13}) \quad (6)$$

Based on above principle, the procedure to find the best maintenance itinerary implemented by M-BAB is demonstrated in Algorithm 1.

**Algorithm 1** Using Modified Brand and Bound method to find the shortest route

```

1: procedure OPTIMAL ROUTE WITH MINIMUM COST
2:    $D = \{D_{jm}\}$  ▷ Distance matrix
3:    $Route = \{0\}$  ▷ Maintenance itinerary is started at global maintenance center (0)
4:    $SoS = \{0, 1, 2, \dots, n\}$  ▷ Set of all visited sites
5:    $CSS = \{S_0\}$  ▷ Complete solution set
6:    $St = 1$  ▷ Declare condition to execute
7:    $q = 0$  ▷ Root node
8:   while  $St < 2$  do
9:     for  $m = SoS - (Route \cap SoS)$  do
10:       $q = q + 1$ 
11:       $VN_{0 \rightarrow q} = \{j | j \in Route\}$  ▷ Visited nodes
12:       $SJ = SoS - Route \setminus \{1, \dots, (end - 1)\} - \{m\}$ 
13:       $LBTC_q = LBTC_{Function}(L, VN_{0 \rightarrow q}, Route, SJ)$ 
14:       $RouteS_q = \{Route, m\}$ 
15:     end for
16:      $[Index, \min Cost] = \min(LBTC)$ 
17:      $CSS = \{CSS, S_{Index}\}$ 
18:      $Route = RouteS_q$  ▷ Update the route
19:     if  $length(Route) = n + 1$  then
20:        $Route_{final} = \{Route, 0\}$ 
21:        $LBTC_{final} = LBTC_{index}$ 
22:        $St = St + 1$ 
23:     end if
24:   end while
25: end procedure
    
```

### B. Efficient evolutionary algorithm (GA – BAB)

The purpose of this section is to develop an efficient evolutionary algorithm to optimize maintenance routing and scheduling in a group. The principle of this algorithm is described as follows:

- **Step 1: Coding.** This step aims at defining the way to introduce a grouping structure in GA-BAB. A grouping structure is here represented by an array  $GS$  in which elements of  $GS$ , denoted  $EGS_p(i^z)$ , is defined:  $EGS_p(i^z) = k$  if  $z_{th}$  maintenance activity of component  $i$  is in group  $k$ . For instance, The coding of a

$$GS = \begin{array}{|c|c|c|c|c|} \hline 1 & 1 & 2 & 3 & 2 \\ \hline \end{array}$$

PM activities    1<sup>1</sup>   2<sup>1</sup>   3<sup>1</sup>   4<sup>1</sup>   1<sup>2</sup>

Fig. 2. Encoding for a grouping structure

grouping structure containing 3 groups  $G_1 = \{1^1, 2^1\}$ ,  $G_2 = \{3^1, 1^2\}$ ,  $G_3 = \{4^1\}$  is shown in Figure 2.

- **Step 2: Generating a population of grouping structures.** GA-BAB creates randomly an initial population of grouping structures. It should be noted that the maximum number of groups in a grouping structure is equal to length of array  $GS$  (denoted  $length(GS)$ ).
- **Step 3: Optimization at group level.** The outstanding of our proposal algorithm than traditional genetic algorithm is that we integrate an optimization process at group level into classical genetic algorithm. Before evaluating economic profit of grouping structure  $EPS$ , economic profit of its each group at group level  $EPG$  should be determined. To do this, for a group  $G_k$ , the M-BAB algorithm is applied to find the optimal departure time ( $t_{G_k}^*$ ) and optimal maintenance itinerary ( $I_{G_k}^*$ ) with the highest grouping economic profit. Based on selected maintenance itinerary and departure time, the performance of a grouping structure can be evaluated at next step.
- **Step 4: Evaluating the performance of grouping structure by fitness function.** The performance of a grouping structure in the population is assessed by its total profit economic  $EPS$ . Based on the grouping structure as well as all optimal itineraries and departure times of its groups, the grouping economic profit can be evaluated as follows:

$$EPS(GS) = \sum_{G_k \in GS} EPG(I_{G_k}^*, t_{G_k}^*) \quad (7)$$

- **Step 5: Elitism.** The two best grouping structures of the current population are directly copied to the next generation in order to protect them from the high level of disruption.
- **Step 6: Crossover.** Crossover is performed to combine a pair of parent grouping structures to generate better grouping structures. To do this, two PM activities are firstly randomly chosen as the crossover points. And then, the elements between these points of the selected parent grouping structures are exchanged (see Figure 3A). The probability that the crossover is done for a pair of grouping structures is around 80%.

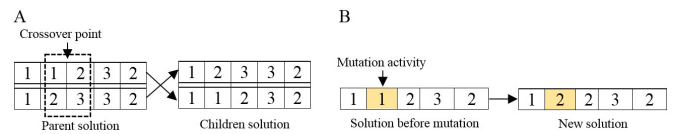


Fig. 3. Example of crossover and mutation operators.

- **Step 7: Mutation.** Mutation helps to prevent GA-BAB from capturing local optima. For each selected grouping

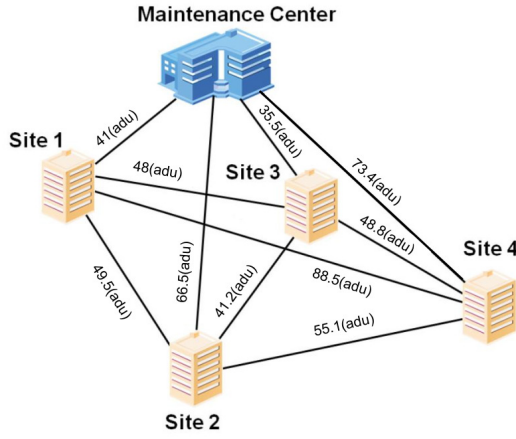


Fig. 4. The typical geographically dispersed production system

structure, a maintenance activity of a group is randomly selected and then moved to another group (see Figure 3B). Mutation probability should be small to prevent GA from random search. It is usually chosen from 1% to 5%.

- *Step 8: New generation.* After implementing the steps 4,5,6, a new population including the best grouping structure of previous population and new grouping structure due to elitism, crossover and mutation is generated. The new generation is evaluated in the next iteration until stop criteria is triggered.
- *Step 9: Stopping.* GA-BAB will be stopped when the maximum number of generations is reached.

Based on above principle, the GAB is simulated based on Algorithm 2.

#### IV. NUMERICAL STUDIES

Our proposed algorithm (GA-BAB) was coded using Matlab language and run on a PC with an Chip (Intel (R) Core(TM) i5-4590) and Ram (4G). In order to assess the efficiency level of our algorithm, we carried out several tests. Firstly, the algorithm is applied to find grouping maintenance plan for a typical GDPS. Secondly, the proposed algorithm is compared to other algorithm (using both exhaustive search and genetic algorithm) in term of computation time to assess its performance. The computation time of simulation is the elapsed time in Matlab expressed in seconds (s).

##### A. Algorithm execution

This subsection executes the proposed algorithm for a GDPS containing 4 sites and one maintenance center as Figure 4. Date of 12 components is shown in Table I and the labor cost rates  $R_{lb}$  are fixed at 100, 200, 400 for required levels of repair team skills 1, 2, 3 respectively. We take  $R^{tr} = 30$ ,  $R_i^{dt} = 80$  and  $S_i^0 = 300$ . It should be noticed that, in this study, all parameters are given in arbitrary units, i.e., arbitrary time unit (atu), arbitrary distance unit (adu) or arbitrary cost unit(acu).

Based on this data, if components are maintained individually, total maintenance cost is equal to 234576.7 (acu). The

**Algorithm 2** Using combination algorithm (GAB) to find the best maintenance plan

```

1: procedure GAB ALGORITHM
2:    $sp = \text{Size of Population}$ 
3:   for  $g=1 \rightarrow sp$  do
4:      $GS_g = \{EGS(i^z) | EGS(i^z) = k\}$   $\triangleright$  Coding
5:   end for
6:    $P = \{GS_g\}$   $\triangleright$  Generating initial population
7:    $iter = 1$ 
8:    $Stop = \text{Number of Iteration}$ 
9:   while  $iter \leq stop$  do  $iter = iter + 1$ ;
10:    for  $g = 1 \rightarrow sp$  do
11:      for  $k = 1 \rightarrow \max\{GS_g\}$  do
12:         $G_k = \{GS(i^z) = k\}$ 
13:         $(t_{G_k}^*, I_{G_k}^*) = \arg \min EPG_{G_k}(t_{G_k}, I_{G_k})$   $\triangleright$ 
        Optimization at group level
14:         $EPG_k^{max} = EPG_k(t_{G_k}^*, I_{G_k}^*)$ 
15:      end for
16:       $EPS_g = \sum_{k=1}^{\max\{GS_g\}} EPG_k^{max}$ 
17:    end for
18:     $g^* = \arg \max \{EPS_g | g = 0 \rightarrow sp\}$ 
19:     $EPS_g^{max} = EPS_{g^*}$   $\triangleright$  Economic profit evaluation
20:    if  $iter = stop$  then  $\triangleright$  Stop condition
21:       $Display(GS_{g^*}, I_{G_k}^*, t_{G_k}^*)$ 
22:    else
23:       $GS_1 = GS_{g^*}$   $\triangleright$  Elitism
24:      for  $g = 2 \rightarrow sp$  do  $\triangleright$  Crossover
25:         $Par1 = \text{random}\{GS_q | q = 2 \rightarrow sp\}$ 
26:         $Par2 = \text{random}\{GS_q | q = 2 \rightarrow sp\}$ 
27:         $ChildrenGS = \text{crossover}(Par1, Par2)$ 
28:         $GS_g = \text{ChildGS}$ 
29:      end for
30:      for  $mg = 1 \rightarrow \text{round}(sp/20)$  do  $\triangleright$  Mutation
31:         $g = \text{random}(3 \rightarrow \text{length}(P))$ 
32:         $p = \text{random}(1 \rightarrow \text{length}(GS_g))$ 
33:         $EGS_p(i^z) = \text{random}(1 \rightarrow \max(GS_g))$ 
34:      end for
35:       $NGS = \{GS_g | g = 1 \rightarrow sp\}$   $\triangleright$  New
        generation
36:    end if
37:  end while
38: end procedure

```

travel distance and transportation cost are equal to 1286.4 (adu) and 38592 (acu). While, using the GA-BAB leads to a grouping maintenance plan shown in Table II.

The proposed maintenance strategy helps to reduce 65.81% travel distance of the maintenance team as well as the travel cost. The reduction of travel distance is very important since it is not only meaningful from economic point of view, but also sustainable one (reduce energy consumption, travel-related risks, environmental negative impacts). Moreover, the grouping maintenance helps to save up to 10.17% total maintenance cost when compare to the individual one.



TABLE I  
DATA OF 12 COMPONENTS

Components	$\lambda_i$	$\beta_i$	$C_i^{sp}$	$C_i^{sc}$	$w_i^p$	$t_i^e$	$l_i$
1	5394	2.2	1045	35	5	706	1
2	6715	2.41	1826	62	8	2365	2
3	4244	1.85	3225	105	12	2508	3
4	5394	2.2	1045	35	5	4052	1
5	6715	2.41	1826	62	8	3279	2
6	4244	1.85	3225	105	12	4194	3
7	5394	2.2	1045	35	5	1589	1
8	6715	2.41	1826	62	8	2633	2
9	4244	1.85	3225	105	12	3904	3
10	5394	2.2	1045	35	5	2531	1
11	6715	2.41	1826	62	8	1022	2
12	4244	1.85	3225	105	12	3294	3

TABLE II  
GROUPING MAINTENANCE PLAN

Group $k$	Components	$t_{G_k}^*$	$L_{G_k}$	$S_{G_k}^{tr}$	$EPG_k$
1	1,2,4,5,7,8,10,11	16849	291.4	8742	17698
2	3,6,9,12	26068	291.4	8742	6176

### B. Performance study of GA-BAB

The aim of this section is to illustrate performance of the proposed algorithm. Then, this performance is assessed by comparing to the performance provided by GA-ES algorithm simulated based on Genetic Algorithm and Exhaustive Search method. The GA-ES process is the same as GA-BAB, except that the Exhaustive Search method (ES) is applied to search the optimal itinerary instead of Branch and Bound method. The Exhaustive Search method aiming to find the optimal itinerary is referred in [14]

1) *Performance of M-BAB*: The objective of this subsection is to assess performance of M-BAB and ES when they are used to find an optimal maintenance plan at group level. Assume that maintenance itinerary of group  $G_k$  has to travel to all sites. When the number of sites rises, the number of maintenance itineraries and computation time also increase with exponential level. The results are shown in Table III and Figure 5. Both M-BAB and ES give the identical results in terms of economic profit and optimal itineraries because they are exact algorithm. However the number of search itineraries ( $N$ ) and computation time ( $CT$ ) of algorithms are different.

From an overall perspective, the number of itineraries and computation time can significantly reduce by using M-BAB. However, when the number of sites is small (3 or 4 sites shown in Table III), the computation times provided by M-BAB and ES is not much different. Indeed, although M-BAB gives the less number of maintenance itineraries than ES, it takes more time to calculate the lower bound function. In case of more sites, performance of M-BAB is significantly better than ES when both the number of itineraries and computation time are reduced sharply as shown in Figure 5.

2) *Performance of GA-BAB and GA-ES*: The objective of this subsection is to assess performance of GA-BAB and GA-ES when they are used to find an optimal maintenance

TABLE III  
COMPARISON BETWEEN M-BAB AND ES

Sites		3	4	5	6	7	8
M-BAB	N	4	17	69	201	3258	24195
	CT (s)	0.394	0.413	0.769	1.989	54.896	363.366
ES	N	6	24	120	720	5040	40320
	CT (s)	0.396	0.519	1.042	5.834	117.977	874.247

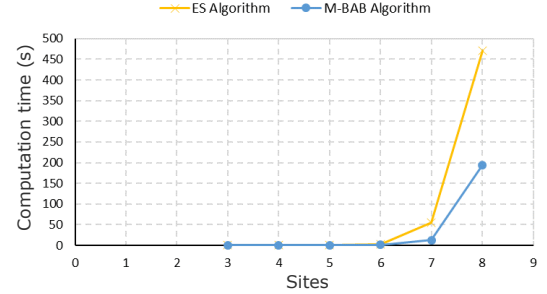


Fig. 5. Comparison of execution time between M-BAB and ES

plan at grouping structure. This work is very difficult because after each iteration, next generation provided by GA-BAB and GA-ES is different due to the random of crossover and mutation. The comparison between two algorithms only can carry out when after each iteration, next generation provided by GA-BAB and GA-ES is identical. For this reason, we do an adjustment procedure in order that next generation provided by GA-BAB and GA-ES can be identical. Assume that initial population of two algorithm is identical, both GA-BAB and GA-ES execute in 10 iterations in cases of the various number of sites. By that way, both algorithms give the identical maintenance plan, however, the computation time has difference. The results of computation time are shown in Table IV and Figure 6.

From the given results, the computation times provided

TABLE IV  
COMPARISON OF COMPUTATION TIME BETWEEN GA-BAB AND GA-ES

Sites	3	4	5	6	7	8
CT (s)						
GA-BAB	1.75	8.42	21.54	55.78	2359.59	21856.19
GA-ES	1.64	4.28	11.15	27.48	1097.85	8720.84

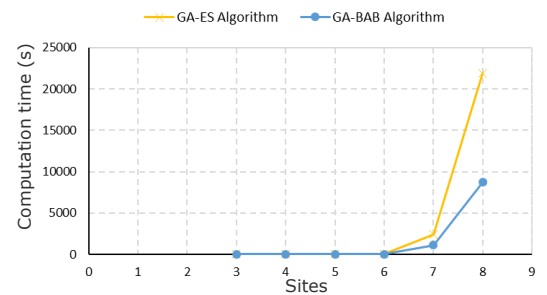


Fig. 6. Comparison of execution time between GA-BAB and GA-ES

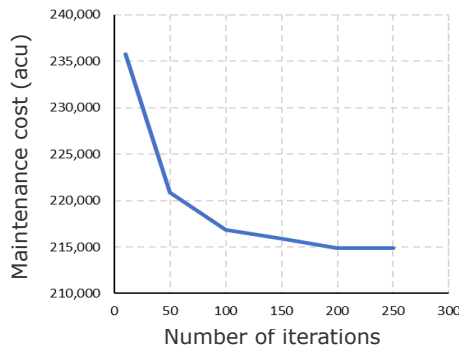


Fig. 7. The convergence curve of GA-BAB algorithm

by GA-BAB and GA-ES are not much different when the number of sites is small. However, in case of more sites, performance of GA-BAB is significantly better than GA-ES when computation time can be reduced by more than 60 %. The convergence curve of the proposed optimization algorithm is shown in Figure 7. The curve will converge when the number of iterations is greater than 200.

## V. CONCLUSION

We have proposed a GA-BAB algorithm for a grouping maintenance planning problem in GDPS context. This algorithm decomposes the problem into optimization problem at group level and grouping structure level. To solve maintenance optimization planning at group level, M-BAB algorithm is proposed to find the shortest maintenance itinerary with an assignment of the technicians to the required tasks. M-BAB enumerates, but constantly tries to rule out parts of the search space that cannot contain the best solution. By that way, search space and computation time reduce significantly. In addition, M-BAB algorithm also determines the departure time of a maintenance team at global maintenance center to minimize penalty cost due to grouping maintenance. For the grouping structure level, the proposed algorithm (GA-BAB) coordinates genetic algorithm and branch and bound to find optimal maintenance plan. The GA-BAB creates a randomly population including grouping structures in which groups of these structures are optimized maintenance scheduling by using M-BAB. Based on phases of GA including elitism, selection, crossover and mutation, GA-BAB give the best maintenance plan.

The performance of M-BAB and GA-BAB has been compared against modified ES and GA-ES algorithm. The proposed algorithms appear to be promising in solving grouping maintenance planning problems. Computational experiments were performed for the proposed M-BAB and GA-BAB with varies of the number of sites, showing difference computation time compared to ES and GA-ES approaches. Experimental results demonstrate the computation time of M-BAB and GA-BAB reducing significantly when the number of sites increase. In fact, the computational experiments and statistical analysis

indicate that the M-BAB and GA-BAB outperform ES and GA-ES in terms of computation time. In the future work, to improve the performance of optimization algorithms, other metaheuristic methods such as simulated annealing or Tabu search can be considered.

## ACKNOWLEDGMENT

This work was supported by The University of Danang, University of Science and Technology, code number of Project: T2020-02-34

## REFERENCES

- [1] F. Camci. Maintenance scheduling of geographically distributed assets with prognostics information. *European Journal of Operational Research*, 245(2):506–516, 2015.
- [2] S. H. Chung, F. T. Chan, and H. K. Chan. A modified genetic algorithm approach for scheduling of perfect maintenance in distributed production scheduling. *Engineering Applications of Artificial Intelligence*, 22(7):1005–1014, 2009.
- [3] P. Do, H. C. Vu, A. Barros, and C. Bérenguer. Maintenance grouping for multi-component systems with availability constraints and limited maintenance teams. *Reliability Engineering & System Safety*, 142:56–67, 2015.
- [4] M. Gendreau, G. Laporte, and F. Semet. A branch-and-cut algorithm for the undirected selective traveling salesman problem. *Networks: An International Journal*, 32(4):263–273, 1998.
- [5] Z. Hameed and K. Wang. Development of optimal maintenance strategies for offshore wind turbine by using artificial neural network. *Wind Engineering*, 36(3):353–364, 2012.
- [6] J. Holland. An introductory analysis with applications to biology, control, and artificial intelligence. *Adaptation in Natural and Artificial Systems. First Edition, The University of Michigan, USA*, 1975.
- [7] C. M. F. Lapa, C. M. N. Pereira, and M. P. de Barros. A model for preventive maintenance planning by genetic algorithms based in cost and reliability. *Reliability Engineering & System Safety*, 91(2):233–240, 2006.
- [8] G. Laporte. The traveling salesman problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59(2):231–247, 1992.
- [9] G. Laporte, Y. Nobert, and S. Taillefer. A branch-and-bound algorithm for the asymmetrical distance-constrained vehicle routing problem. *Mathematical Modelling*, 9(12):857–868, 1987.
- [10] E. López-Santana, R. Akhavan-Tabatabaei, L. Dieulle, N. Labadie, and A. L. Medaglia. On the combined maintenance and routing optimization problem. *Reliability Engineering & System Safety*, 145:199–214, 2016.
- [11] K. S. Moghaddam and J. S. Usher. Preventive maintenance and replacement scheduling for repairable and maintainable systems using dynamic programming. *Computers & Industrial Engineering*, 60(4):654–665, 2011.
- [12] J. S. Srai, M. Kumar, G. Graham, W. Phillips, J. Tooze, S. Ford, P. Beecher, B. Raj, M. Gregory, M. K. Tiwari, et al. Distributed manufacturing: scope, challenges and opportunities. *International Journal of Production Research*, 54(23):6917–6935, 2016.
- [13] A. Volkanovski, B. Mavko, T. Boševski, A. Čauševski, and M. Čepin. Genetic algorithm optimisation of the maintenance scheduling of generating units in a power system. *Reliability Engineering & System Safety*, 93(6):779–789, 2008.
- [14] S. Vukmirović, Z. Čapko, and A. Babić. Model of using the exhaustive search algorithm in solving of traveling salesman problem (tsp) on the example of the transport network optimization of primorje-gorski kotar county (pgc). In *7th International OFEL Conference on Governance, Management and Entrepreneurship: Embracing Diversity in Organizations. April 5th-6th, 2019, Dubrovnik, Croatia*, pages 391–401. Zagreb: Governance Research and Development Centre (CIRU), 2019.

# Using Dual-use Electronic Lectures in E-learning: An Empirical Study of Teaching and Learning Mathematics at Vietnamese High Schools

Bui Anh Tuan

*Department of Mathematics Education  
Teachers College, Can Tho University  
Can Tho City, Vietnam  
batuan@ctu.edu.vn*

Tran Thi Thu Thao\*

*Department of Mathematics Education  
Teachers College, Can Tho University  
Can Tho City, Vietnam  
Luu Huu Phuoc High School  
Can Tho City, Vietnam  
thuthao.maths.edu@gmail.com*

Nguyen Ngoc Phuong Anh

*Department of Mathematics Education  
Teachers College, Can Tho University  
Can Tho City, Vietnam  
anhb1700003@student.ctu.edu.vn*

Le Thanh Dien

*Department of Mathematics Education  
Teachers College, Can Tho University  
Can Tho City, Vietnam  
lethanhdienst@gmail.com*

**Abstract**— E-learning is a form of teaching with several advantages, especially in the context of disease outbreaks such as COVID-19. In terms of teaching at High Schools, E-learning lectures are utilized both for direct classroom teaching and for online teaching via the Internet. This article presents a dual-use electronic lecture design model (DUEL) with iSpring Suite, which can be flexibly used for both classroom and online learning goals. An empirical research on DUEL is conducted in Mathematics at high school. After the experiment, a survey that evaluating the effectiveness of using DUEL has sent to learners. The feedback results showed that the first step of using DUEL was received quite positively and had developed to their full potential in fluctuations of today's teaching context.

**Keywords**—E-learning, Blended learning, Dual-use electronic lecture, Mathematics education, High school.

## I. INTRODUCTION

In the digital age, E-learning is an effective knowledge-providing strategy with many supporting tools and has proven effective over time [1]. As we know, E-learning is seen as a knowledge management system that allows many educational institutions to take advantage of learning anytime and anywhere [2], so that learners can personalize their learning when self-control over content, sequence, time and speed of knowledge acquisition [3]. On the other hand, E-learning provides content, knowledge base, facilitates access to learning-temporary resources called E-learning materials. In addition, it also supports online multimedia exchange with media tools like Facebook, YouTube, Google Classroom,... creates interactive discussion environment through the Internet [4].

E-learning has gradually become a widespread learning method in higher education institutions, but there are some limitations, especially at High School level [5]. The participation of student is utmost important and significant for online learning, from which observational and applied learning behavior indicators can be learned. This online learning method requires higher student self-discipline than a traditional classroom, especially for students who are not

disciplined and have limited IT skills, they cannot complete the required tasks [6]. According to Zhang et al. [7], E-learning efficiency is influenced by numerous factors such as the learning context, technological factors, characteristics, psychology of learners,... Moreover, the dropout rate is high if there is no teacher reminder. Based on Juutinen & Saariluoma [8] and Juutinen [9], this digital approach also creates some psychological and emotional obstacles for both teachers and learners in interactive processes.

During outbreaks such as COVID-19, many schools are closed, UNESCO recommends the use of distance learning programs and open educational applications. According to Zayapragassarazan [10], online lectures are focused on to ensure the continuity of the teaching and learning process. In addition, Mahalakshmi & Radha [11] pointed out that, E-learning methodology demonstrates its role with functions such as creating virtual classes, sharing resources, checking and evaluating online,... On the other hand, the author also made a point: “COVID 19 to be negative but it is POSITIVE towards e-learning”.

As we know, the “School's Out, But Class's On” & “Suspending Classes Without Stopping Learning” are policies proposed by the Chinese government with the goal of not disrupting studying in this country due to the effects of viruses [12], [13]. Teaching online is proposed by many authors to cope with the pandemic COVID-19 such as Verawardina et al. [14], Basilaia & Kvavadze [15]. Thus it can be said that online learning is one of the effective solutions during the COVID-19 pandemic.

In terms of practicality, depending on the context, E-learning has a certain role. This study proposes the design of E-learning materials as dual-use electronic lectures, can be used to teach directly in the classroom, and can be used to teach online in cases of force majeure such as coronavirus outbreak. To initially evaluate the effectiveness of E-learning and Blended learning, we have conducted a study on teaching Mathematics in the context of High School. This study was

conducted from January 2018 to January 2020, based on iSpring Suite software and 3D tools.

The outline of this paper is structured as follows. Section 2 presents about literature review. Section 3 illustrates methodology and some of results. Conclusion is offered in the last section.

## II. LITERATURE REVIEW

### A. E-learning and Blended learning

E-learning is often defined based on several aspects such as the use of technology to support learning, approaches to learning resources, learning processes and interactions, and issues of improvement of traditional educational models. E-learning is the use of electronic media for learning purposes to replace face-to-face meetings [16] by distance online courses [17], learners access documents using a computer, phone or mobile device [18].

In addition, E-learning is to provide multimedia learning materials such as text, graphics, animation, audio or video [3], provided via email, Internet and the platform of World Wide Web (WWW) [19], the learner is received via electronic media [20]. E-learning is an educational method in which communication is computerized, pedagogical interaction between learners and materials, learners and learners, learners and teachers through online Web as chat, video call, and so on [21].

E-learning is the use of new multimedia digital technologies and the Internet to improve the quality of learning [22], create conditions to receive knowledge anytime and anywhere [23], improve the learners' knowledge and skills with cooperation and remote exchange [24], without having to worry about changing time and space [25]. Thus it can be seen that E-learning is considered a new learning method based on information technology in which multimedia material is designed with the purpose of easy access, multidimensional interaction through wireless devices, communication and computer networks to provide knowledge, skills, improve the quality of learning and teaching. The following diagrams are showing the relationship between E-learning in educational models (see Fig. 1).

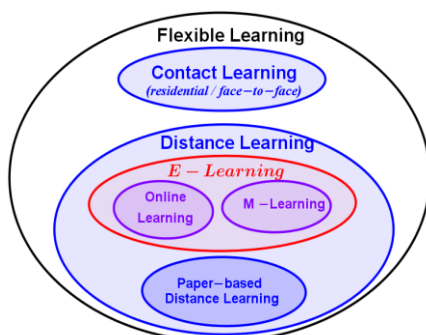


Fig. 1. The subsets of flexible learning by Brown [26].

It can be observed from Fig. 1 that the diagram of the relationship between forms of learning by Brown [26], E-learning is the type of distance learning with either Online or M-learning options.

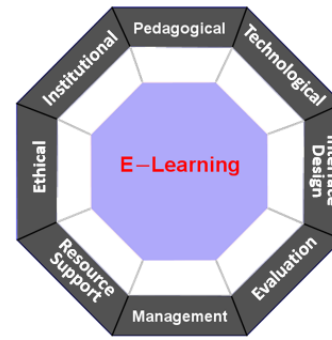


Fig. 2. The E-learning framework by Khan [27]

As we can see from Fig. 2 that, E-learning is a combination of eight elements: Institutional, Management, Technological, Pedagogical, Ethical, Interface design, Resource support, Evaluation. It has been seen that these elements are analogous to the way a traditional classroom is organized. The question to be asked is: "can E-learning completely replace face-to-face teaching method of a traditional classroom?"

Let's take a brief look at the learning process in a traditional classroom. According to Hasebrook et al. [28], this environment creates opportunities for developing relationships between learners and learners and between teachers and learners (so-called face-to-face), communication skills, group discussion, argument to come up with ideas well done when preserving the nuance of speech [29].

Hence it can be seen that E-learning has many advantages but still exists more limited than traditional classes. Because of the lack of immediate response, it is delayed through intermediaries such as email, chat, after a few hours or days [30]. Furthermore, motivations of student should be encouraged by the teacher but due to the absence of face-to-face contact, this factor causes depression and anxiety for learners. Moreover, Bell & Federman [31] mentioned to the problem of cheating while studying online.

As we know, there have been numerous research authors comparing the advantages and limitations of the two above-mentioned learning methods, see e.g.; Bates & Poole [32], Zhang et al. [7], Rovai et al. [33], Kanninen [34] and Caravias [35]. Most of them conclude that they need to combine their advantages. One of the proposals is the Blended learning model (see Fig. 3).

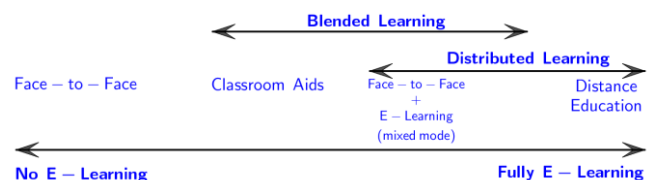


Fig. 3. The continuum of technology-based learning by Bates & Poole [32]

Blended learning combines two processes: face-to-face learning in traditional classroom and online learning together [36], brings many advantages in the future, reduces lecture time with open resources [37], expands learning environment in space and time [37], [39]. Thus it has been seen that Blended learning is a guide to the traditional learning environment supported by technology elements of E-learning



to create visual, vivid and positive learning motivation. This is clearly shown by the following diagram (see Fig. 4).



Fig. 4. The Challenge of finding Blends that take advantage of the strengths of each environment and avoid the weaknesses by Graham [40]

According to Graham [40], Blended learning process is converted based on the proportion of content distributed online depending on the conditions and the actual situation. For instance, face-to-face has 0% online, Blended learning has 30%-79% online and Online Learning has over 80% online. In addition, Picciano [41] provides a diagram showing the transition of the combination from traditional environment to Online as follows (see Fig. 5):

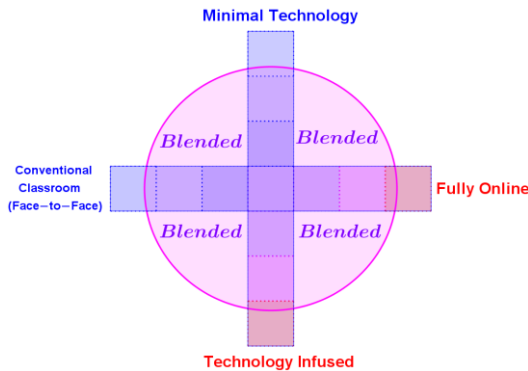


Fig. 5. Broad Conceptualization of Blended learning by Picciano [41]

Online learning content includes extra readings, notes, pictures, graphs or extended practical issues [41]. For example, teachers direct students in class as usual and create more online discussions to address projects outside the classroom. Therefore, the main concern is how this transition happens in the classroom. We propose the use of online learning materials (can be called electronic lectures) to convey the core knowledge in a face-to-face and online process. The primary goal is to support the expansion of knowledge for students with unlimited criteria of space and time.

### B. Dual-use Electronic Lectures and iSpring

Barker and Tan [42] pointed out that the use of electronic lessons with integrated multimedia (multimedia), multi-dimensional interaction with the purpose of supporting teaching and learning activities. The main goal is to improve ordinary presentations into an open, fast-growing learning resource. Nevertheless, electronic lectures used for classroom teaching and online teaching are often independent. Teachers must take a lot of effort to convert electronic lectures into appropriate teaching forms.

In the field of technology, the term "Dual-use Technology" is often used to refer to technologies used for civil and military fields [43], [44]. In education, dual-use electronic lessons are used in traditional classroom teaching, transmitted to learners during self-study at home, Online learning before or after the

lesson. In research, Dual-use Electronic Lectures are understood as follows: "Dual-use Electronic Lectures (DUEL) is a learning resource used in combination in two environments, Online learning and Traditional learning, capable of switching between online and offline forms. The primary purpose is to suit many different learning goals, in diverse contexts of society".

To design DUEL iSpring Suite software is a preferred option with many outstanding features. It is an add-in integrated with Microsoft PowerPoint (PPT), has all the features of E-learning, has an easy-to-see and easy-to-use interface and a 14-day trial at the address:

<https://www.ispringsolutions.com/>

## III. RESEARCH METHODOLOGY & RESULTS

### A. Research Methodology

To establish a DUEL design model, collect feedback when using DUEL in teaching, the research is conducted according to the following process (see Fig. 6):

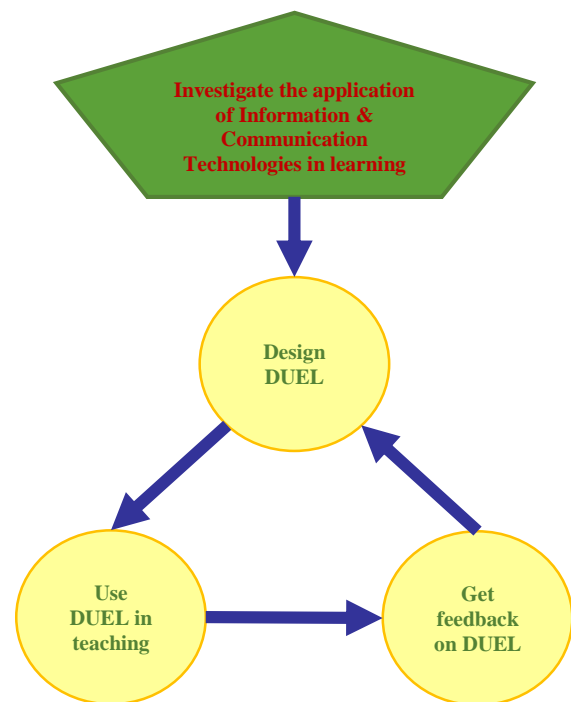


Fig. 6. The Procedure of Establishing DUEL (PED).

In step 1: "Survey on the application of Information & Communication Technologies in learning" was conducted by survey on 35 students of High School of Can Tho University, Can Tho city, Vietnam. It is important that students employ the tools and methods to find knowledge for the reports when requested, thereby analyzing the feasibility and necessary conditions to plan the design of DUEL. Results and analysis are described in section B.

Next to step 2: "Design DUEL". The study conducted to design 5 DUEL corresponding to 5 lessons of chapter "Vectors and perpendicular relations in space", Geometry of grade 11. The primary purpose is to extend the time and enhance the consolidation of knowledge before and after class, DUEL has materials that have been prepared in digital format, combining PPT with some features of iSpring Suite such as Quiz, Interactions, Simulations, Youtube, Video, Games,... It has also been published in various forms (Wed,



CD, LMS, Video,...) and the interface is suitable for many devices (Laptop, Mobile phone,...). This study publishes DUEL in the form of Web to provide online materials and CD format for offline use. DUELS are stored in the Google Drive cloud at the address:

[https://drive.google.com/drive/folders/1AImTTbcST3cflXl6gV5f\\_s5A3HvAybp7?usp=sharing](https://drive.google.com/drive/folders/1AImTTbcST3cflXl6gV5f_s5A3HvAybp7?usp=sharing)

The process in step 3 "Using DUEL in teaching" is conducted as follows: Lesson vector in space (DUEL 01) and Two perpendicular lines (DUEL 02) Offline format, in which the teacher teaches knowledge and guides students to use DUEL. Lesson perpendicular to the plane (DUEL 03) published in 2 forms: Offline and Online, after class, students can manipulate themselves with DUEL 03 at home to revise knowledge, perform a Quiz and this result will give feedback to teachers via Gmail.

For perpendicular two-sided lessons (DUEL 04) sent to students 1 week in advance for students to study, go to group discussion class, re-emphasize the core knowledge, especially with DUEL 03, students are required to pass 80% of the final test to access DUEL 05. DUEL 05 provides the knowledge of the lesson about Distance and revision of the whole chapter, teachers support online students and 2 classes in class. The process of increasing the rate of self-study with DUEL is based on [40], [41].

Finally, the research "collecting feedback on DUEL" with the survey. This method is based on evaluation of educational products by survey and experiment [45]. Results and analysis are described in section C.

#### B. ICT survey results

In this study, the ICT's questionnaire will give students 5 questions in order to assess their abilities in applying information technology throughout their knowledge accomplishments.

In **question A1**: Please indicate your daily usage for the media (gmail, social networks, phones, laptops, internet, etc.) of yourself. Within 35 answers, 23 percent of them are "always", 63 percent are "often", 11 percent are "sometimes" and 3 percent are "seldom" and none of them is "never".

In **question A2**: Which means do you use to access the internet or search, exploit information and materials for learning? (Can choose multiple options)? With the sample size of 35 observations ( $N = 35$ ), 74 percent of them use mobile phone, 71 percent use desktop or laptop and 14 percent use tablet. The most notable thing is that almost every student own an instrument which can access the Internet to facilitate their studies.

As shown from the survey above, most students nowadays are available to use the Internet quite frequently by using their mobile phone, desktop or tablet. This seems to be a good standard to connect their studies process with the information technology platform, notably DUEL. Assumed that before applying the DUEL basis into teaching procedure, the infrastructures and students' abilities in ICT are limited, we should introduce and have some guidelines to help them get used to and have a better experience while using this standard.

In **question A3**: Do you agree with teaching in combination with information and communication technology? With the infrastructure and the student's skill,

from **question A3**, they are recently surveyed in the level of agreement of learning using the ICT-aid foundation. The result was 11 percent of them are strongly agree, 57 percent of agreement, 32 percent of consideration.

In **question A4**: If you were asked to prepare a report, group exercise and need to present to the class, which editing tool would you use? (Can choose multiple options).

We collected the result with 97 percent of Microsoft PowerPoint, 2 percent of Latex, 3 percent of Lecture Maker, 6 percent of GeoGebra, 0 percent of Violet and some other options. In this question, Microsoft PowerPoint with the highest portion seems to be a good signal. As presented in the previous section, Microsoft PowerPoint is the basic core in designing DUEL lessons. Students who frequently access and choose MsPP throughout their studies will find it easy to deal with and DUEL is anticipated to bring the optimal result.

Continue with **question A5**: If you were asked to prepare a report, group exercise and need to present to the class, which editing tool would you use? (Can choose multiple options). This question concentrate on the abilities/skills in searching for Mathematic materials.

For ease of data processing, we numbered the advantages from (i) to (viii) as follows:

- (i). Read textbooks or reference books.
- (ii). Communicate by meeting directly with friends and teachers.
- (iii). Paper documents provided by teachers.
- (iv). View electronic reference book online.
- (v). Talk with friends and teachers via online chat or video call.
- (vi). Web material provided by Web sites.
- (vii). Join forums or social networking sites, online study groups
- (viii). Others

Advantages	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
Quantity	12	19	25	9	16	20	14	1
Proportion	34 %	54 %	71 %	26 %	46 %	57 %	40 %	3 %

With the above survey results, realize that the survey team knows how to use ICT to search for learning materials, have skills to work with PPT (the foundation of DUEL). Therefore, it has been seen that the equipment conditions are good, and there is a willingness to study with ICT. From that, it can be concluded that DUEL based on PPT design is familiar and easy to manipulate for students and using DUEL in teaching is feasible and can bring higher learning efficiency.

#### C. Feedbacks from using DUEL in teaching

After using DUEL, we conduct a survey to find the effect that DUEL have on psychology as well as student results through 5 question from **B1** to **B5**

**Question B1**: What is your attitude when studying Mathematics with DUEL? Given 5 levels of preference: strongly like, like, neither dislike nor like, dislike and strongly dislike. The highest portion was really like and like with 77 percent (68 of them was like), 23 percent still consider in applying DUEL into learning process, particularly Mathematics.

**Question B2**: In your opinion, is studying Math through DUEL attractive? Collected result shows that 6 percent is strongly interested, 63 percent of them see DUEL interested, 31 percent still choose neutral and none of the two left options

**Question B3:** How would you rate your comprehension for your Math lesson with DUEL? (Select the larger number, the better understand the article). The ordinal scale have been coded by following with the understanding levels to calculate the sample mean.

Score	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$	$x_1 = 5$
Frequency	$n_1 = 0$	$n_2 = 2$	$n_3 = 10$	$n_4 = 15$	$n_5 = 8$

Mean is calculated as followed:

$$\bar{X} = \frac{\sum_{i=1}^5 x_i n_i}{N} \approx 3.83$$

This survey result with the mean GPA equals 3.83 over 5 absolute mark, DUEL allows student to repeat knowledge again and again over time and control each student input through testing and no limitation of time and space. In short, it is may be due to DUEL's "window" style sequential test design makes the comprehension rate high at the experimental students' group.

**Question B4:** In your opinion, what are the advantages of studying Math in combination with DUEL? (Can choose multiple options).

To ease data processing, we numbered the advantages from (I) to (X) as follows:

- (I). Can self-study anytime and anywhere.
- (II). Can review the learning content repeatedly.
- (III). There is a knowledge test after each topic or chapter.
- (IV). Meet the individuality in learning (learning options and content).
- (V). Free reference materials are provided.
- (VI). Be entertained through games during the lecture.
- (VII). Attractive and vivid images, videos, materials attached lecture.
- (VIII). Help students understand lessons more, content more accurately.
- (IX). Help learners be more active and positive in learning.
- (X). Other comments.

In **question B4**, we have listed out some advantages of DUEL and continue to survey in order to evaluate the DUEL's design quality, thus tender renovation (if needed).

Advantages	(I)	(II)	(III)	(IV)	(V)
Quantity	24	32	16	14	19
Percentage	69%	91%	46%	40%	54%
Advantages	(VI)	(VII)	(VIII)	(IX)	(X)
Quantity	12	25	12	19	7
Percentage	34%	71%	34%	54%	20%

From the result, student realize some significant advantages like: "Can review the learning content repeatedly"; "Be entertained through games during the lecture"; "Meet the individuality in learning" is the difference compared with traditional lessons.

Moreover, student also prefer some other opinions:

- (X<sub>1</sub>): "Student find it more interesting in study"  
 (X<sub>2</sub>): "Attractive images and graphics"  
 (X<sub>3</sub>): "Higher study efficiency than normal"  
 (X<sub>4</sub>): "Student are able to review before and after learning on their on"  
 (X<sub>5</sub>): "Knowledge available at anytime, you just need a smart phone"  
 (X<sub>6</sub>): "You need to finish the older lesson' tests in order to move on to another one, this is a nice procedure to optimal your study efficiency"

With opinions listed, we affirm that most of student who have approach DUEL can understand the advantages as well as what it them during their study process

**Question B5.** When studying Math in combination with DUEL, what do you think you need to improve?

With this question, some suggestions are raised to improve on DUEL's publication quality, one of them was "to combine 3D tools and imitation" in order to illustrate game' graphics, models as well as mathematic graph of function

Notwithstanding the application of DUEL in general teaching and teaching of Mathematics is a new method, it is accepted by the experimental group (expressed by students' level of interest, comprehension, engaging level of DUEL). In question B4, the advantages that DUEL has in the teaching process are clearly shown. It is obvious that the contributions of the experimental group at B5 will help the future improvement and development of DUEL

#### IV. CONCLUSION

This research again shows the great important and significant role of E-learning in numerous different contexts of education, especially in High Schools. Dual-use lectures (DUEL) with flexible transition have been showing the superiority in saving lecture time, creating more excitement for students. It can be seen that it is particularly highly adaptable in a variety of particular teaching contexts, such as the COVID-19 pandemic.

From an application perspective, in addition to Mathematics, DUEL can be applied and executed in several other High School subjects such as Physics, History, English, and so on. From a technology perspective, besides iSpring Suite, DUEL can be deployed on numerous other software and applications; on E-learning systems like Moodle or Web 2.0 platforms. In the near future, with the growing development of mobile technology, DUEL promises a new potential when integrated with different types of devices.

Regarding the establishment process of DUEL (PED), the experiment shows that the initial step is suitable for the High School context, with young people who have easy access to information technology in learning. The empirical survey results also indicate the versatility of PED: not only can it be applied in Mathematics but it also has the potential to apply in many other subjects at High School. Moreover, in the near future, PEDs can be performed in higher education environment, especially in countries with young population such as ASEAN region.

#### REFERENCES

- [1] Rosenberg, M. J., & Foshay, R. (2002). E-learning: Strategies for delivering knowledge in the digital age. *Performance Improvement*, 41(5), 50-51.
- [2] Shehabat, I. M., & Mahdi, S. A. (2009, April). E-learning & its Impact to the Educational System in the Arab World. In *2009 International Conference on Information Management & Engineering* (pp. 220-225). IEEE.
- [3] Ruiz, J. G., Mintzer, M. J., & Leipzig, R. M. (2006). The impact of E-learning in medical education. *Academic medicine*, 81(3), 207-212.
- [4] Clark, R. C., & Mayer, R. E. (2016). *E-learning & the science of instruction: Proven guidelines for consumers & designers of multimedia learning*. John Wiley & Sons.
- [5] Mahanta, D., & Ahmed, M. (2012). E-learning objectives, methodologies, tools & its limitation. *International Journal of Innovative Technology & Exploring Engineering (IJITEE)*, 2(1), 46-51.

- [6] Wong, D. (2007). A critical literature review on E-learning limitations. *Journal for the Advancement of Science & Arts*, 2(1), 55-62.
- [7] Zhang, D., Zhao, J. L., Zhou, L., & Nunamaker Jr, J. F. (2004). Can E-learning replace classroom learning?. *Communications of the ACM*, 47(5), 75-79.
- [8] Juutinen, S., & Saariluoma, P. (2010). Emotional Obstacles for E-learning--A User Psychological Analysis. *European journal of Open, Distance & E-learning*.
- [9] Juutinen, S. (2011). *Emotional Obstacles of E-learning* (No. 145). University of Jyväskylä.
- [10] Zayapragassaran, Z. (2020). COVID-19: Strategies for Online Engagement of Remote Learners. *F1000Research*, 9.
- [11] Mahalakshmi, K., & Radha, R. COVID 19: A MASSIVE EXPOSURE TOWARDS WEB BASED LEARNING.
- [12] Zhang, W., Wang, Y., Yang, L., & Wang, C. (2020). Suspending classes without stopping learning: China's education emergency management policy in the COVID-19 Outbreak.
- [13] Zhou, L., Wu, S., Zhou, M., & Li, F. (2020). 'School's Out, But Class' On', The Largest Online Education in the World Today: Taking China's Practical Exploration During The COVID-19 Epidemic Prevention & Control As an Example. *But Class' On', The Largest Online Education in the World Today: Taking China's Practical Exploration During The COVID-19 Epidemic Prevention & Control As an Example* (March 15, 2020).
- [14] Verawardina, U., Asnur, L., Lubis, A. L., Hendriyani, Y., Ramadhani, D., Dewi, I. P., ... & Sriwahyuni, T. (2020). Reviewing Online E-learning Facing the Covid-19 Outbreak. *Journal of Talent Development & Excellence*, 12(3s), 385-392.
- [15] Basilaia, G., & Kvavadze, D. (2020). Transition to online education in schools during a SARS-CoV-2 coronavirus (COVID-19) p&emic in Georgia. *Pedagogical Research*, 5(4), 1-9.
- [16] Guri-Rosenblit, S. (2005). 'Distance education' and 'E-learning': Not the same thing. *Higher education*, 49(4), 467-493.
- [17] Marquès, P. (2006). Definición del E-learning. *Retrieved from pangea.org/peremarques*.
- [18] Governors State University, Center for Online Learning and Teaching. (2008). *E-learning glossary*.
- [19] Gunasekaran, A., McNeil, R. D., & Shaul, D. (2002). E-learning: research and applications. *Industrial and commercial training*.
- [20] Koohang, A., & Harman, K. (2005). Open source: A metaphor for E-learning. *Informing Science*, 8.
- [21] Videgaray, M. G. (2007). Evaluación de la reacción de alumnos y docentes en un modelo mixto de aprendizaje para educación superior.
- [22] Alonso, F., López, G., Manrique, D., & Viñes, J. M. (2005). An instructional model for web-based e-learning education with a Blended learning process approach. *British Journal of educational technology*, 36(2), 217-235.
- [23] Okiki, O. C. (2011). Information Communication Technology Support for an ELearning Enviroment at the University of Lagos, Nigeria.
- [24] Bhuasiri, W., Xaymoungkhoun, O., Zo, H., Rho, J. J., & Ciganek, A. P. (2012). Critical success factors for E-learning in developing countries: A comparative analysis between ICT experts and faculty. *Computers & Education*, 58(2), 843-855.
- [25] Mbarek, R., & Zaddem, F. (2013). The examination of factors affecting E-learning effectiveness. *International Journal of Innovation and Applied Studies*, 2(4), 423-435.
- [26] Brown, T. H. (2003, June). The role of m-learning in the future of E-learning in Africa. In *21st ICDE World Conference* (Vol. 110, pp. 122-137).
- [27] Khan, B. H. (2003). The global E-learning framework. *STRIDE*, 42.
- [28] Hasebrook, J., Herrmann, W., & Rudolph, D. (2003). *Perspectives for European E-learning businesses: Markets, technologies and strategies* (No. 47). Office for Official Publications of the European Communities.
- [29] Hameed, S., Badii, A., & Cullen, A. J. (2008, May). Effective E-learning integration with traditional learning in a Blended learning environment. In *European and Mediterranean Conference on Information Systems* (pp. 25-26).
- [30] Blass, E., & Davis, A. (2003). Building on solid foundations: establishing criteria for E-learning development. *Journal of further and higher education*, 27(3), 227-245.
- [31] Bell, B. S., & Federman, J. E. (2013). E-learning in postsecondary education. *The future of children*, 165-185.
- [32] Bates, T., & Poole, G. (2003). Effective teaching with technology in higher education: Foundations for success.
- [33] Rovai, A., Ponton, M., Wighting, M., & Baker, J. (2007). A comparative analysis of student motivation in traditional classroom and E-learning courses. *International Journal on E-learning*, 6(3), 413-432.
- [34] Kanninen, E. (2009). Learning styles and E-learning. *Tampere: Tampere University of Technology*, 1, 5-29.
- [35] Caravias, V. (2015). Literature review in conceptions and approaches to teaching using blended learning. In *Curriculum Design and Classroom Management: Concepts, Methodologies, Tools, and Applications* (pp. 1-22). IGI Global.
- [36] Thorne, K. (2003). *Blended learning: how to integrate online & traditional learning*. Kogan Page Publishers.
- [37] Watson, J. (2008). Blended Learning: The Convergence of Online and Face-to-Face Education. Promising Practices in Online Learning. *North American Council for Online Learning*.
- [38] Hou, S. (2012). Construction and Application of the Network Class in Blending Learning Method. *IERI Procedia*, 2, 561-564.
- [39] Tayebinik, M., & Puteh, M. (2013). Blended learning or E-learning?. *Tayebinik, M., & Puteh, M.(2012). Blended learning or E-learning*, 103-110.
- [40] Graham, C. R. (2006). Blended learning systems. *The handbook of blended learning: Global perspectives, local designs*, 3-21.
- [41] Picciano, A. G. (2006). Blended learning: Implications for growth and access. *Journal of asynchronous learning networks*, 10(3), 95-102.
- [42] Barker, P., & Tan, C. M. (1997). Making a case for electronic lectures. *Innovations in education and training international*, 34(1), 11-16.
- [43] Atlas, R. M., & Dando, M. (2006). The dual-use dilemma for the life sciences: perspectives, conundrums, and global solutions. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 4(3), 276-286.
- [44] Forge, J. (2010). A note on the definition of "dual use". *Science and Engineering Ethics*, 16(1), 111-118.
- [45] Mertens, D. M. (2014). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Sage publications.

# Biometric Image Recognition For Secure Authentication Based on FPGA : A survey

Huu Q Tran

<sup>a</sup>*Department of Electronics Technology  
Industrial University of Ho Chi Minh City  
Ho Chi Minh City, Vietnam*

<sup>b</sup>*Department of Electrical and electronic Engineering  
Ho Chi Minh City University of Technical Education  
Ho Chi Minh City, Vietnam  
quyhuutran@gmail.com*

Van Thai Nguyen

*Department of Electrical and electronic Engineering  
Ho Chi Minh City University of Technical Education  
Ho Chi Minh City, Vietnam  
vanthainguyen@gmail.com*

**Abstract**—Field Programmable Gate Arrays (FPGA) based biometric image recognition is one of the highly secure authentication technologies. It bases on personal unique features of each person. The employment of FPGA provides specific reprogrammable hardware solution that can be properly exploited to obtain a reconfigurable recognition system. This paper presents a survey of biometric image recognition that relies on FPGAs. These biometric recognition approaches include face recognition and iris recognition. This paper is a brief review of the concept and structure of biometric recognition systems. Furthermore, several recognition algorithms are also surveyed in this work. In addition, a summary of the state of the art of each recognition type is also provided as a background for findings in biometric recognition systems.

**Index Terms**—biometric recognition, face recognition, iris recognition, FPGA

## I. INTRODUCTION

Biometric recognitions have played an important role in secure authentication applications. Biometric authentication is an identification technology based on the unique physiological and behavioral traits, which involves fingerprints, ears, face, voiceprint, keystroke dynamics, iris, and the shape of the body [1]. Among these biometric technologies, the face and iris recognition are two new and the most accurate recognition techniques in the current technologies. Compared to traditional authentication methods such as passwords, PIN, and identification cards, they show the advantages of highly accurate and reliable human identification mechanisms. While the traditional authentication methods can be stolen, lost, forgotten or shared easily, biometric-based authentication (i.e. human authentication) is a recognition using physiological and behavioral features that are greatly distinguished among people to verify the identity of a person. The biometric recognitions include fingerprint, face, voice, iris. These characteristics are constant throughout its lifetime and unique to each person [2].

Currently, FPGAs have been considered as a platform of the choice for reconfigurable hardware implementation of real-time image processing applications [3]. FPGAs are programmable devices by the end user [4]. The structure of FPGAs consists of an array of uncommitted elements that

can be reconfigured and upgraded easily according to a users specifications. The design with very high performance can be achieved using FPGAs due to their high density and high performance as compared to digital signal processing (DSP) systems [5]. A high degree of parallelism and orders of magnitude speedup are the superior features of FPGAs [6].

Motivated by the aforementioned analysis, this paper presents an overview of biometric recognition including face recognition and iris recognition based on FPGA. The contributions of this work are summarized into three main ideas.

- A brief introduction about FPGA and its state of the art in biometric recognition is presented.
- Parameters for evaluating of a biometric recognition system are provided in this work.
- Two biometric recognition techniques, namely face recognition and iris recognition, are investigated to provide an overview of current biometric recognition technologies. Concepts, system models, the most common face recognition and iris recognition algorithms are discussed and analyzed in this work. In addition, the state of the art of two types of biometric recognition techniques are also presented in this work.

## II. FPGA OVERVIEW

As mentioned in Section I, FPGAs enable the end-user to reconfigure their structure for obtaining the final logic system. The structures of the FPGAs are inherently parallel and include a huge number of registers, high-speed memory and storage interfaces and embedded memory blocks. An FPGA contains an array of programmed or interconnected elements so that the users can generate the logic structure according to their specification in a virtually limitless number of ways. In FPGAs, these elements are logic blocks which are interconnected to form a bi-dimensional array. Each logic block contains look-up tables (LUTs) to store Boolean functions. The LTUs are constructed over simple memories [7]. In principle, by using the appropriate configuration, FPGAs can implement any combinational or sequential circuits. Hence, FPGAs show a combination between the

hardware-based speed of ASICs and the flexibility of general-purpose processors. FPGAs can provide ultra-low power platforms for optimizing energy consumption [7]. Moreover, FPGA can parallelly process data which results in high speed [8]. FPGA can be commonly programmed via Hardware Description Languages such as Verilog [9] or VHDL [10].

Table 1 presents the state of the art of FPGA based biometric recognition systems.

TABLE I: The state of the art of FPGA based biometric recognition

Ref	Year	FPGA version	FPGA based biometric recognition type
[11]	2004	Altera FLEX10K FPGA	Iris recognition
[12]	2007	Virtex IV- xc4vlx25-11, Xilinx	Fingerprint recognition
[13]	2008	Spartan 3 XC3S2000, Xilinx	Minutiae extraction in Fingerprint Recognition
[14]	2010	Altera Stratix II FPGA	Finger Vein Biometrics
[15]	2011	Virtex-4 XC4VFX12 FPGA, Xilinx ML403	Face recognition
[16]	2018	prototyping FPGA Zed-Board, Xilinx	Face Recognition
[17]	2020	DE-10 nano Cyclone V SOC	Iris recognition

### III. FPGA BASED BIOMETRIC IMAGE RECOGNITION

#### A. Performance parameters

Basic methods are most commonly used to measure the performance of biometric technologies such as Performance Histogram, False Reject Rate (FRR), False Accept Rate (FAR), Crossover Error Rate (CER), Failure to Acquire/Failure to Enrol (FTA/FTE) and D and ROC Plots [18], [19].

- FRR (False Reject Rate): indicates the number of times in which an authorized user is wrongly refused to access to the protected system. False rejects are often caused by incorrect positioning of the hand or finger, smudges on a finger scanner, incorrect alignment of the retina or iris.

- FAR (False Accept Rate): indicates the number of times in which an unauthorized user is accepted and thus wrongly accepted to the protected system, therefore enabling a security breach.

- CER (Crossover Error Rate)/ EER (Equal Error Rate): is sometimes used to identify system accuracy. It shows where the FRR and the FAR are equal. The lower the CER, the higher the accuracy of the system.

- FTA (Failure to Acquire)/ FTE (Failure to Enrol): happens when data from the biometric feature is not collected enough by the capture equipment, e.g. sunlight shining on an iris scanning capture device or a finger pressing down too hard on the platen.

#### B. Face recognition

Face Recognition is one of the biometric technologies which is used for the identification of a person among the pool of images. It is most widely used in security systems all over the world. This approach matches the unique characteristics of a human face with the database. Thus, the accuracy of a face recognition system strongly depends on the algorithm for feature extraction. Goldstein, Harmon and Lesk are the first group at Bell Telephone Laboratories which proposed the implementation of the problem of face recognition [20]. 255 faces were used for analysis and assessment of face recognition in which 21 subjective features such as hair color, lip thickness were considered in order to automate face recognition. Eigenfaces for face recognition were developed by Kirby and Sirovich in 1988 showing that less than 100 values are needed to approximately normalize and suitably align face image. The advantages of face recognition include non-invasive sensing, lower cost, portable dimension, tracking of people with a longer distance at domestic or public places and public integration on embedded systems for users access identity [21]. However, the face recognition system is not the most reliable and efficient compared with other biometric systems such as fingerprint, eye, iris recognition systems [22].

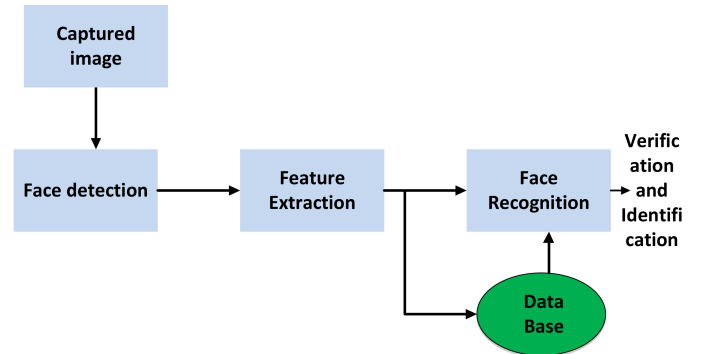


Fig. 1: An illustration of a basic face recognition system.

A face recognition system includes three basic steps as follows (1) face detection, (2) feature extraction and (3) face recognition [23]. A general block diagram of a face recognition system is illustrated in figure 1.

There are several approaches which have been studied in face detection and feature extraction for face recognition. Among these approaches, Principal Component Analysis (PCA), i.e. Eigenfaces, is one of the most known global face recognition algorithms in face recognition. This method bases on finding the eigenvectors ( i.e. principal directions ) of the covariance matrix of the multidimensional data to de-correlate data and to feature similarities and differences in the faces. Eigenfaces are the features of the human face. Figure 2 illustrates a diagram of the PCA method.

Table 2 introduces the state of the art of several algorithms employed in face detection and feature extraction of face recognition systems.



TABLE II: State of the art of algorithms for face detection and feature extraction in face recognition

Ref	Face detection	Feature extraction	FPGA implementation
[24]	Local Binary Pattern (LBP) and Gabor Filter	Local Binary Pattern (LBP) and Gabor Filter	–
[25]	Scale Invariant Feature Transform algorithm	Scale Invariant Feature Transform algorithm	Zynq 7000 AP SoC ZC 702 Evaluation Kit -Board
[26]	Eigenface/Principal Component Analysis (PCA) algorithm	Eigenface/Principal Component Analysis (PCA) algorithm	DE2 Altera board, a Cyclotrone II FPGA
[27]	Viola-Jones face detection algorithm	–	DE2-115 evaluation board
[28]	Local Ternary Pattern (LTP)	–	Xilinx Virtex-7 FPGA
[29]	Fast Fourier Transform (FFT)	FFT	–
[30]	–	The Haar-Like Features	–

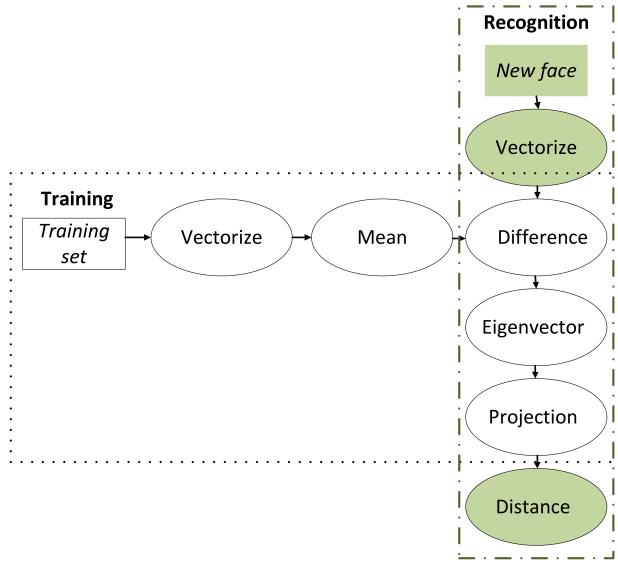


Fig. 2: An illustration of PCA method.

### C. Iris Recognition

Iris recognition is one of the most accurate and high confidence for authentication and verification methods that have been used today. The features of the iris of each person are special, unchanged during the lifetime and impossible to manipulate with it. The iris recognition can be applied in national ID schemes, border control, access control, Surveillance, Law enforcement, Service industry, Military, Robotics [31].

Basically, a human eye is divided into three parts: the pupil, the iris and the sclera as shown in Figure 3. The black circular hole with dark black color in the center of the eye is defined the pupil. The colored ring-shaped object positioned between the white part of the eye and the pupil is defined the iris. The iris is covered by two boundaries within the eye. The pupillary boundary is determined by a line separating the black pupil in the center of the eye from the colored iris. The limbic boundary is determined by a line separating the colored iris from the sclera white part of the eye. The iris is one of the most unique structures in the human body since it is created by many detailed features of the glandular fossa, pigment spots, wrinkles, etc [32]. Furthermore, the development of the iris enters a relatively stable period at the 8-month fetus stage [32]. As a result, the stability, unchangeable and uniqueness

nature of the iris is the background for identical recognitions. The white part at the periphery of the eyeball is defined the sclera.

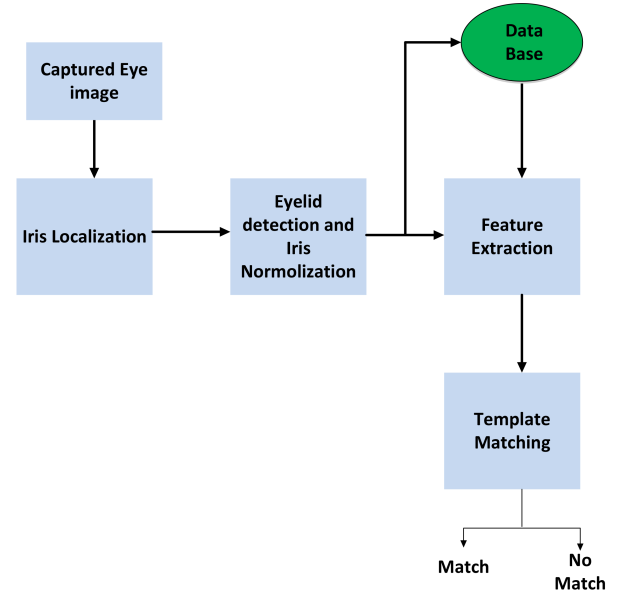


Fig. 4: An illustration of a basic iris recognition system.

The iris recognition indicates the comparison of the similarity between iris image features to examine a persons identity. To perform iris recognition, the iris is collected as samples in advance, then constructed an iris database and establish the relationship between iris data and personal information. An iris recognition framework normally consists of five main steps as shown in Figure 4: image acquisition, iris localization, normalization, feature extraction and template matching.

#### a) Normalization

After the successful iris region segmentation from an eye image, the iris region is transformed to obtain fixed dimensions for comparison in the normalization stage. As a result, the characteristic features of two images of the same iris under different conditions are the same at the same spatial locations. It is noted that the pupil region is usually slightly nasal and not always concentric within the iris region [33]. The transformation of the iris region can be performed using Daugmans homogenous Rubber Sheet Model [18] to remap each point of iris to a pair of polar coordinates  $(\theta, r)$ , where

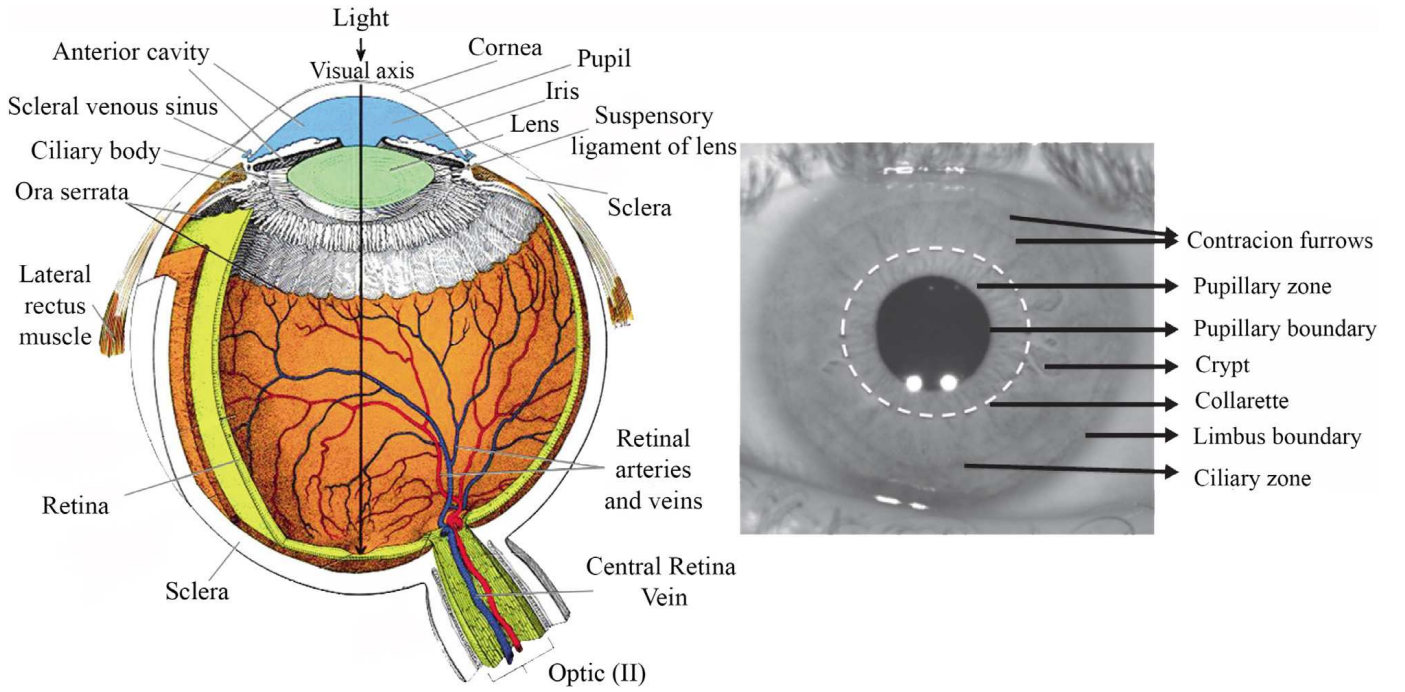


Fig. 3: An illustration of the structure of the iris (a) Inner and (b) Outer [31].

$\theta \in [0, 2\pi]$  and  $r \in [0, 1]$

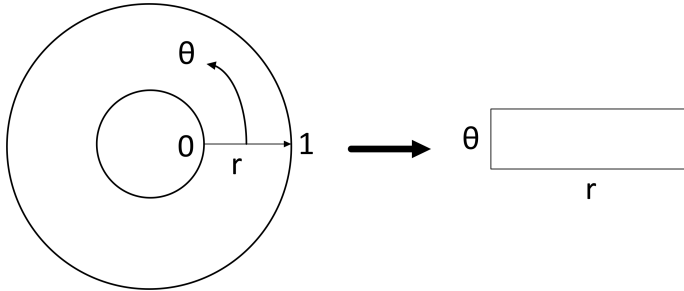


Fig. 5: An illustration of Daugman's rubber sheet model.

Due to the non-concentric feature of the pupil and iris regions, it must be considered when normalizing the doughnut-shaped iris region to achieve a constant radius. The re-mapping of the iris region image can be expressed by [34], [35]

$$I(x(r, \theta), y(r, \theta)) \rightarrow I(r, \theta) \quad (1)$$

Where  $(x, y)$  is the original Cartesian coordinate system and  $(r, \theta)$  is the corresponding normalized non-concentric polar coordinate system

$x(r, \theta), y(r, \theta)$  are determined by [34], [35]

$$\begin{aligned} x(r, \theta) &= (1 - r)x_p(\theta) + rx_s(\theta) \\ y(r, \theta) &= (1 - r)y_p(\theta) + ry_s(\theta) \end{aligned} \quad (2)$$

Where  $x(r, \theta)$  and  $y(r, \theta)$  are the linear combination of both the set of limbus boundary points along the outer perimeter of the iris bordering the sclera and the set of pupillary

boundary points.  $(x_s, y_s)$  and  $(x_p, y_p)$  are coordinates of iris and pupil boundaries along the  $\theta$  direction.

#### b) Localization

This stage is a crucial stage in the success of any iris recognition system. When the localization of the iris region from an eye image is successful, the next step is to represent each iris image as a graph of a fixed iris region so that it has fixed dimensions for comparisons. However, the iris pattern data will corrupt the generated biometric templates since data that is falsely represented. As a result, the recognition rates are poor. A common method used to localize an iris is known as Hough Transform. It was first reported by Paul Hough in 1962 [36]. It is an algorithm that allows determining parameters in an image within a certain geometrical form such as an ellipse, circle, and line.

The representation of any circle with center  $(x_c, y_c)$  and radius  $r$  is given by [37]

$$r^2 - x_c^2 - y_c^2 = 0 \quad (3)$$

Considering the obtained edge points of an edge map of the image as for the parameters of circles passing through each edge points as  $(x_i, y_i)$ . The Hough transform is expressed by [37],

$$H(x_c, y_c, r) = \sum_{i=1}^n h(x_i, y_i, x_c, y_c, r) \quad (4)$$

$$h(x_i, y_i, x_c, y_c, r) = \begin{cases} 1, & \text{if } g(x_i, y_i, x_c, y_c, r) = 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Where  $g(x_i, y_i, x_c, y_c, r) = (x_i - x_c)^2 + (y_i - y_c)^2 - r^2$

TABLE III: State of the art of algorithms for iris recognition

Ref	Iris localization	Iris normalization	Feature extraction	Matching and database	FPGA implementation
[42]	Hough transform	Mapping, bilinear interpolation, equalizing the gray-level histogram of the iris image	1D DCT	Hamming distances, CASIA	–
[43]	Local Kurtosis	Polar coordinate transformation	A vertical filter and a horizontal filter	Hamming Distance	DE2 with Cyclone-II EP2C35 FPGA chip, Altera
[44]	Hough Transform, accelerating the construction of histograms	–	–	–	a Xilinx Zynq-7000 XC7Z020
[45]	A proposed circle Hough transform (CHT) by using three accumulator arrays	–	–	CITHV4, CIV1	Xilinx's 7 series Zynq FPGA
[46]	Non-iterative method	A polar coordinate transformation, an non-iterative method		CASIA-IrisV3-Lamp, MMU v1.0, ND-IRIS-0405 and NIST ICE 2005	Cyclone IV EP4CE115 FPGA
[47]	Hough transform, active contour method	Active contour method	2D Log-Gabor filter	CASIA interval V3	–

### c) Feature extraction

This process extracts the information from the iris image for classification. The recognition rate of an algorithm to a great extent is determined by the quality of feature extraction. The feature extraction is based on local sharp variations to construct a set of the characteristics of the iris that is known as a sort of transient signals. The process starts by locating the pupil of the eye, which can be performed using any edge detection technique. Only the significant features of the iris must be encoded for the comparison between templates.

### d) Matching and classification

The matching is a process of comparing between two templates to determine their matching. One template is the current template and the other has already been stored in the database. This process is performed until finding one template in the database which matches the current template. This stage commonly utilizes the distance-based approach to match and classify the extracted iris image. Several matching algorithms, e.g. Hamming distance [38] [39], Euclidean distance, weight vector or neural network, are usually employed to denote the distance. Classification is performed by template matching against stored template to verify one to one matching or identify one to many matching

The following presents the Hamming distance as an example of a matching algorithm. The Hamming distance is utilized to measure of how close two templates to each other. The more the Hamming distance closes to zero, the higher the accurate matching. As reported by Daugman in [40], the highest closeness is 0.32. The Hamming distance expression is given by [33], [41]

$$HD = \frac{\|(Tp\_codeA \otimes Tp\_codeB) \cap maskA \cap maskB\|}{\|maskA \cap maskB\|} \quad (6)$$

Where  $\otimes$  is the XOR operator.  $\cap$  is the AND operator.  $\| \cdot \|$  is the norm of the AND'ed mask vectors and of the resultant bit

vector.  $Tp\_codeA$  and  $Tp\_codeB$  are the phase code bit vector of templates A and B, respectively.  $maskA$  and  $maskB$  are the mask bit vectors of the phase code bit vectors, respectively.

Table 3 summarizes the state of the art of several algorithms employed in each phase of iris recognition systems.

## IV. CONCLUSION

This paper has presented an overview of FPGA based biometric recognition in terms of face recognition and iris recognition. An introduction and state of art of FPGA in biometric recognition is discussed and summarized. Two face and iris recognition techniques with several algorithms are provided. Furthermore, two popular algorithms including Hough transform and Hamming distance are analyzed in detail. In addition, several parameters for biometric recognition evaluation are presented. This finding is a basic background for studying on biometric recognition.

## REFERENCES

- [1] Mitra, S., Gofman, M. (Eds.). (2016). Biometrics in a Data Driven World: Trends, Technologies, and Challenges. CRC Press.
- [2] Ross, A., Nandakumar, K., Jain, A. K. (2008). Introduction to multibiometrics. In Handbook of biometrics (pp. 271-292). Springer, Boston, MA
- [3] Mittal, S., Gupta, S., Dasgupta, S. (2008, June). FPGA: An efficient and promising platform for real-time image processing applications. In National Conference On Research and Development In Hardware Systems (CSI-RDHS).
- [4] Brown, S. D., Francis, R. J., Rose, J., Vranesic, Z. G. (2012). Field-programmable gate arrays (Vol. 180). Springer Science and Business Media.
- [5] Memik, S. O., Katsaggelos, A. K., Sarrafzadeh, M. (2003). Analysis and FPGA implementation of image restoration under resource constraints. IEEE transactions on Computers, 52(3), 390-399.
- [6] Ma, J. (2003). Signal and Image processing via Reconfigurable Computing. In Proc. of the First Workshop on Information and Systems Technology (pp. 1-6).
- [7] De La Piedra, A., Braeken, A., Touhafi, A. (2012). Sensor systems based on FPGAs and their applications: A survey. Sensors, 12(9), 12235-12264.
- [8] Rakvic, R. N., Ngo, H., Broussard, R. P., Ives, R. W. (2010). Comparing an FPGA to a Cell for an Image Processing Application. EURASIP Journal on Advances in Signal Processing, 2010(1), 764838.

- [9] Palnitkar, S. (2003). Verilog HDL: a guide to digital design and synthesis (Vol. 1). Prentice Hall Professional.
- [10] Heinkel, U., Glauert, W., Wahl, M. (2000). The VHDL Reference: A Practical Guide to Computer-Aided Integrated Circuit Design (Including VHDL-AMS) with Other. John Wiley and Sons, Inc..
- [11] Mohd-Yasin, F., Tan, A. L., Reaz, M. I. (2004, December). The FPGA prototyping of Iris recognition for biometric identification employing neural network. In Proceedings. The 16th International Conference on Microelectronics, 2004. ICM 2004. (pp. 458-461). IEEE.
- [12] Lindoso, A., Entrena, L., Izquierdo, J. (2007, February). FPGA-based acceleration of fingerprint minutiae matching. In 2007 3rd southern conference on programmable logic (pp. 81-86). IEEE
- [13] Lopez, M., Canto, E. (2008, June). FPGA implementation of a minutiae extraction fingerprint algorithm. In 2008 IEEE International Symposium on Industrial Electronics (pp. 1920-1925). IEEE
- [14] Khalil-Hani, M., Eng, P. C. (2010, October). FPGA-based embedded systolic implementation of finger vein biometrics. In 2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA) (pp. 700-705). IEEE
- [15] Sudha, N., Mohan, A. R., Meher, P. K. (2011). A self-configurable systolic architecture for face recognition system based on principal component neural network. IEEE transactions on circuits and systems for video technology, 21(8), 1071-1084
- [16] Bonny, T., Rabie, T., Hafez, A. A. (2018). Multiple histogram-based face recognition with high speed FPGA implementation. Multimedia Tools and Applications, 77(18), 24269-24288
- [17] Ma, L., Sham, C. W. (2019, December). SoC-FPGA-Based Implementation of Iris Recognition Enhanced by QC-LDPC Codes. In 2019 International Conference on Field-Programmable Technology (ICFPT) (pp. 391-394). IEEE
- [18] Sanderson, S., Erbetta, J. H. (2000). Authentication for secure environments based on iris scanning technology.
- [19] Bolle, R. M., Pankanti, S., Ratha, N. K. (2000, September). Evaluation techniques for biometrics-based authentication systems (FRR). In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000 (Vol. 2, pp. 831-837). IEEE.
- [20] Goldstein, A. J., Harmon, L. D., Lesk, A. B. (1971). Identification of human faces. Proceedings of the IEEE, 59(5), 748-760.
- [21] Gaurav R. Chimote, N.M. Tarbani (2016). A survey paper on authentication system in android phones, International Journal of Scientific Engineering Research.
- [22] Yang, W., Wang, S., Hu, J., Zheng, G., Valli, C. (2018). A fingerprint and finger-vein based cancelable multi-biometric system. Pattern Recognition, 78, 242-251.
- [23] Napolon, T., Alfalou, A. (2017). Pose invariant face recognition: 3D model from single photo. Optics and Lasers in Engineering, 89, 150-161.
- [24] Gorde, S. H., Kumar, M. M., Balramudu, M. P. (2016). An FPGA based Face Recognition System using Gabor and Local Binary Pattern. network, 5(1).
- [25] Peng, J. Q., Liu, Y. H., Lyu, C. Y., Li, Y. H., Zhou, W. G., Fan, K. (2016, June). FPGA-based parallel hardware architecture for SIFT algorithm. In 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR) (pp. 277-282). IEEE.
- [26] Sustersic, T., Vulovic, A., Filipovic, N., Peulic, A. (2016). FPGA implementation of face recognition algorithm. In Pervasive Computing Paradigms for Mental Health (pp. 93-99). Springer, Cham
- [27] Irgens, P., Bader, C., Le, T., Saxena, D., Ababei, C. (2017). An efficient and cost effective FPGA based implementation of the Viola-Jones face detection algorithm. HardwareX, 1, 68-75.
- [28] Byun, J. Y., Jeon, J. W. (2017). FPGA based face detection using local ternary pattern under variant illumination condition. In Advances in Computer Science and Ubiquitous Computing (pp. 365-370). Springer, Singapore.
- [29] Sustersic, T., Peulic, A. (2019). Implementation of face recognition algorithm on field programmable gate array (FPGA). Journal of Circuits, Systems and Computers, 28(08), 1950129
- [30] Alahmadi, A., Qaisar, S. M. (2019, March). Robust Real-time Embedded Face Detection Using Field Programmable Gate Arrays (FPGA). In 2019 Advances in Science and Engineering Technology International Conferences (ASET) (pp. 1-5). IEEE.
- [31] Nguyen, K., Fookes, C., Jillela, R., Sridharan, S., Ross, A. (2017). Long range iris recognition: A survey. Pattern Recognition, 72, 123-143.
- [32] Yi, K., Deng, Q., Yuan, B., Qu, X., Gao, J., Fernandes, T. (2018, November). Iris Recognition and Data Storage on Cloud. In 2018 Asia-Pacific Magnetic Recording Conference (APMRC) (pp. 1-3). IEEE.
- [33] Daugman, J. (2009). How iris recognition works. In The essential guide to image processing (pp. 715-739). Academic Press.
- [34] Chawla, S., Oberoi, A. (2011). A robust algorithm for iris segmentation and normalization using hough transform. Global Journal of Business Management and Information Technology, 1(2), 69-76.
- [35] Daugman, J. (2007). New methods in iris recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 37(5), 1167-1175.
- [36] HOUGH, Paul VC. Method and means for recognizing complex patterns. U.S. Patent No 3,069,654, 1962.
- [37] Farouk, R. M. (2011). Iris recognition based on elastic graph matching and normalization using hough wavelets. Computer Vision and Image Understanding, 115(8), 1239-1244.
- [38] Rai, H., Yadav, A. (2014). A Unified approach for unconstrained off angle Iris Recognition. Expert systems with applications, 41(2), 588-593.
- [39] Hariprasath, S., Mohan, V. (2008, December). Biometric Personal Identification based on Iris recognition using complex wavelet transforms. In 2008 International Conference on Computing, Communication and Networking (pp. 1-5). IEEE.
- [40] Daugman, J. G. (1993). High confidence visual recognition of persons by a test of statistical independence. IEEE transactions on pattern analysis and machine intelligence, 15(11), 1148-1161.
- [41] Omran, S. S., Al-Hilali, A. (2015, September). Using an FPGA to Accelerate Iris Recognition. In 2015 International Conference on Advances in Software, Control and Mechanical Engineering (ICSCME2015).
- [42] Donald M. Monro, Soumyadip Rakshit, and Dexin Zhang (2007), DCT-Based Iris Recognition, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.
- [43] Ryan N. Rakvic, Member, IEEE, Bradley J. Ullis, Randy P. Broussard, Robert W. Ives, Neil Steiner (2009), Parallelizing Iris Recognition, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.
- [44] Orlando, C., Andrea, P., Christophel, M., Xavier, D., Granado, B. (2018). FPGA-Based Real Time Embedded Hough Transform Architecture for Circles Detection. In 2018 Conference on Design and Architectures for Signal and Image Processing (DASIP) (pp. 31-36). IEEE
- [45] Vineet Kumar , Abhijit Asati, Anu Gupta (2018), Memory-efficient architecture of circle Hough transform and its FPGA implementation for iris localisation, IET.
- [46] Tariq M. Khan, Donald G. Bailey, Mohammad A. U. Khan, Yanan Kong (2019), Realtime iris segmentation and its implementation on FPGA, Journal of Real-Time Image Processing.
- [47] Ammour, B., Boubchir, L., Bouden, T., Ramdani, M. (2020). FaceIris Multimodal Biometric Identification System. Electronics, 9(1), 85.

# A novel two-variable model for bending analysis of laminated composite beams

Xuan-Bach Bui

Faculty of Civil Engineering

Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam

Trung-Kien Nguyen

Faculty of Civil Engineering

Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam

Quoc-Cuong Le

Faculty of Engineering - Technology

Thu Dau Mot University

Binh Duong Province, Viet Nam

cuongtd91@gmail.com

T. Truong-Phong Nguyen

Faculty of Civil Engineering

Ho Chi Minh City University of Technology and Education  
Ho Chi Minh City, Viet Nam

**Abstract**—A novel two-variable model for static analysis of laminated composite beams is proposed in this paper. The kinematics of the beam having only two variables are expanded in a hybrid form under polynomial and trigonometric series in thickness and axial directions, respectively. Lagrange's equations are then used to derive characteristic equations of the beams. Numerical results for laminated composite beams are compared with previous studies and are used to investigate the effects of length-to-depth ratio, fibre angles and material anisotropy on the deflection and stresses of laminated composite beams.

**Index Terms**—Laminated composite beams; Bending; Elasticity solution.

## I. INTRODUCTION

Laminated composite materials are fabricated by assembling multiple layers of fibrous materials to achieve the superior engineering properties such as bending stiffness, strength-to-weight ratio and thermal performance. As a result, laminate composite has been widely applied in aerospace engineering, mechanical engineering as well as construction technology. In order to maximise the potential advantage of this multilayered material, numerous studies and computation modelling have been conducted to fine-tune the static and dynamic behaviours of laminated composite beams. Various beam theories have been developed in order to predict accurately their structural responses and capture anisotropy of laminated composite materials. Classical beam theory (CBT) is the simplest one in analyzing responses of laminated composite beams. Nonetheless, this theory underestimates deflections and overestimates natural frequencies of the beams due to neglecting effects of transverse shear deformation. In order to account for this effect, thanks to its simplicity in formulation and programming, the first-order shear deformation beam theory (FSBT) is commonly used by researchers and commercial softwares for the analysis of laminated composite beams ([1], [2], [3], [4], [5]). However, in this theory, the inadequate distribution of

transverse shear stress in the beam thickness requires a shear correction factor to calculate the shear force. This adverse in practice could be overcome by using higher-order deformation beam theory (HSBT) ([6], [7], [8], [9], [10], [11], [12], [13], [14], [15]) or Quasi-3D beam theory (Quasi-3D) ([16], [17], [18], [19], [20], [21], [22]) owing to the higher-order variation of axial displacement or both axial and transverse displacements, respectively. In such approach, stresses of the beam can be directly computed from constitutive equations without shear coefficient requirement. Many higher-order shear deformation theories have been developed with different approaches in which its kinematics could be expressed in terms of polynomial ([23], [24], [25], [26], [27]), trigonometric ([28], [29], [30], [31], [32], [33], [34]), exponential ones ([35], [36]), hyperbolic ([37], [38], [39]) and hybrid higher-order shear functions ([40], [41]). A literature review shows that a vast number of researches on development HSBT and Quasi-3D have been developed, however the accuracy of these theories strictly depends on the choice of shear functions and number of variables defining the problem. The development of new beam theories as well as suitable solution methods is a complicated problem and needs to study further.

The objective of this paper is to develop a bi-directional elasticity solution for static analysis of laminated composite beams. Based on the elasticity equations, the proposed theory only requires two unknowns in which the axial and transverse displacements are approximated in series terms in its two in-plane directions for different boundary conditions and Lagrange's equations are used to derive characteristic equations. Numerical results are presented to investigate the effects of length-to-depth ratio, fibre angle and material anisotropy on the deflections and stresses of laminated composite beams.

## II. THEORETICAL FORMULATION

Considering a laminated composite beam with rectangular section  $b \times h$  and length  $L$ , the beam is composed of  $n$  layers of orthotropic materials.

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant No. 107.02-2018.312.



### A. Kinematic, strain and stress

Denoting  $u$  and  $w$  are axial and transverse displacements at location  $(x, z)$  of the beam. The linear displacement-strain relations of the beam are given by:

$$\epsilon_x = \frac{\partial u}{\partial x} \quad (1a)$$

$$\epsilon_z = \frac{\partial w}{\partial z} \quad (1b)$$

$$\gamma_{xz} = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \quad (1c)$$

Based on an assumption of the plan stress in the plane  $(x, z)$  of the beam, i.e.  $\sigma_y = \sigma_{yz} = \sigma_{xy} = 0$ , the elastic constitutive equation at the  $k^{th}$ -layer in the global coordinate system is expressed by:

$$\begin{Bmatrix} \sigma_x \\ \sigma_z \\ \sigma_{xz} \end{Bmatrix} = \begin{bmatrix} \bar{C}_{11} & \bar{C}_{13} & 0 \\ \bar{C}_{13} & \bar{C}_{33} & 0 \\ 0 & 0 & \bar{C}_{55} \end{bmatrix} \begin{Bmatrix} \epsilon_x \\ \epsilon_z \\ \gamma_{xz} \end{Bmatrix} \quad (2)$$

where  $\bar{C}_{11}$ ,  $\bar{C}_{13}$  and  $\bar{C}_{55}$  are the reduced in-plane and out-of-plane elastic stiffness coefficients of the laminated composite beam in the global coordinates (see [22] for more details).

### B. Energy formulation

The total static energy  $\Pi$  of the beam under an external transverse loading comprises the strain energy  $\mathcal{U}$  and work done by the external load  $\mathcal{V}$ . The strain energy  $\mathcal{U}$  of the beam is given by:

$$\begin{aligned} \mathcal{U} &= \frac{1}{2} \int_V (\sigma_x \epsilon_x + \sigma_z \epsilon_z + \sigma_{xz} \gamma_{xz}) dV \\ &= \frac{1}{2} \int_V \left\{ \bar{C}_{11} \left( \frac{\partial u}{\partial x} \right)^2 + 2\bar{C}_{13} \frac{\partial u}{\partial x} \frac{\partial w}{\partial z} + \bar{C}_{33} \left( \frac{\partial w}{\partial z} \right)^2 \right. \\ &\quad \left. + \bar{C}_{55} \left[ \left( \frac{\partial u}{\partial z} \right)^2 + 2\frac{\partial u}{\partial z} \frac{\partial w}{\partial x} + \left( \frac{\partial w}{\partial x} \right)^2 \right] \right\} dV \end{aligned} \quad (3)$$

The work done by a transverse load  $q$  at the bottom surface of the beam is given by:

$$\mathcal{V} = - \int_0^L q w dx \quad (4)$$

The total energy of the beam is therefore expressed by:

$$\begin{aligned} \Pi &= \frac{1}{2} \int_V \left\{ \bar{C}_{11} \left( \frac{\partial u}{\partial x} \right)^2 + 2\bar{C}_{13} \frac{\partial u}{\partial x} \frac{\partial w}{\partial z} + \bar{C}_{33} \left( \frac{\partial w}{\partial z} \right)^2 \right. \\ &\quad \left. + \bar{C}_{55} \left[ \left( \frac{\partial u}{\partial z} \right)^2 + 2\frac{\partial u}{\partial z} \frac{\partial w}{\partial x} + \left( \frac{\partial w}{\partial x} \right)^2 \right] \right\} dV \\ &\quad - \int_0^L q w dx \end{aligned} \quad (5)$$

### C. Bi-directional Ritz solution

Based on the Ritz method, the axial and transverse displacements at location  $(x, z)$  of the beam can be generally approximated in the following forms:

$$u(x, z, t) = \sum_{r=1}^R \sum_{s=1}^S \psi_{rs}(x, z) u_{rs} \quad (6a)$$

$$w(x, z, t) = \sum_{r=1}^R \sum_{s=1}^S \varphi_{rs}(x, z) w_{rs} \quad (6b)$$

where  $u_{rs}, w_{rs}$  are unknown displacement values to be determined;  $\psi_{rs}(x, z)$ ,  $\varphi_{rs}(x, z)$  are the bi-directional shape functions which are composed of admissible hybrid exponential-trigonometric function in the  $x$ -axis and polynomial function in the  $z$ -axis as follows:

$$\begin{aligned} \text{S-S} : \quad \psi_{rs}(x, z) &= \cos \frac{\pi x}{L} e^{-rx/L} z^{s-1} \\ \varphi_{rs}(x, z) &= \sin \frac{\pi x}{L} e^{-rx/L} z^{s-1} \end{aligned} \quad (7a)$$

$$\begin{aligned} \text{C-F} : \quad \psi_{rs}(x, z) &= \sin \frac{\pi x}{2L} e^{-rx/L} z^{s-1} \\ \varphi_{rs}(x, z) &= \left( 1 - \cos \frac{\pi x}{2L} \right) e^{-rx/L} z^{s-1} \end{aligned} \quad (7b)$$

$$\begin{aligned} \text{C-C} : \quad \psi_{rs}(x, z) &= \sin \frac{\pi x}{L} e^{-rx/L} z^{s-1} \\ \varphi_{rs}(x, z) &= \sin^2 \frac{\pi x}{L} e^{-rx/L} z^{s-1} \end{aligned} \quad (7c)$$

It is noted that the shape functions in Eqs. (7) satisfy kinetic boundary conditions of the beams (S-S: simply supported beams, C-F: clamped-free beams, C-C: clamped-clamped beams). The characteristic equations of the beam can be obtained by substituting Eqs. (6) into Eq. (5) and using Lagrange's equations:

$$\frac{\partial \Pi}{\partial q_{rs}} = 0 \quad (8)$$

where  $q_{rs} = (u_{rs}, w_{rs})$  are displacement vector of the beam. The characteristic equations of the beam are obtained as follows:

$$\begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} \\ {}^T \mathbf{K}^{12} & \mathbf{K}^{22} \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{w} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{F} \end{Bmatrix} \quad (9)$$

where the components of stiffness matrix  $\mathbf{K}$  and load vector

$F$  are defined as follows:

$$\begin{aligned}
K_{rspq}^{11} &= \int_0^L \int_{-h/2}^{h/2} \bar{C}_{11} \frac{\partial \psi_{rs}}{\partial x} \frac{\partial \psi_{pq}}{\partial x} b dx dz \\
&+ \int_0^L \int_{-h/2}^{h/2} \bar{C}_{55} \frac{\partial \psi_{rs}}{\partial z} \frac{\partial \psi_{pq}}{\partial z} b dx dz, \\
K_{rspq}^{12} &= \int_0^L \int_{-h/2}^{h/2} \bar{C}_{13} \frac{\partial \psi_{rs}}{\partial x} \frac{\partial \varphi_{pq}}{\partial z} b dx dz \\
&+ \int_0^L \int_{-h/2}^{h/2} \bar{C}_{55} \frac{\partial \psi_{rs}}{\partial z} \frac{\partial \varphi_{pq}}{\partial x} b dx dz, \\
K_{rspq}^{22} &= \int_0^L \int_{-h/2}^{h/2} \bar{C}_{33} \frac{\partial \varphi_{rs}}{\partial z} \frac{\partial \varphi_{pq}}{\partial z} b dx dz \\
&+ \int_0^L \int_{-h/2}^{h/2} \bar{C}_{55} \frac{\partial \varphi_{rs}}{\partial x} \frac{\partial \varphi_{pq}}{\partial x} b dx dz, \\
F_{rs} &= \int_0^L q \varphi_{rs} dx
\end{aligned} \quad (10)$$

### III. NUMERICAL EXAMPLES

A range of numerical examples are performed in this section to verify the efficiency of the present theory with different boundary conditions. The laminated composite beam is subjected to a uniformly distributed load applied on the bottom surface and in the  $z$ -direction of the beam. Laminates are assumed to have equal thicknesses and are made of the same orthotropic materials whose properties are given as follows: material I ( $E_1/E_2 = 40$ ,  $E_2 = E_3$ ,  $G_{12} = G_{13} = 0.6E_2$ ,  $G_{23} = 0.5E_2$ ,  $\nu_{12} = \nu_{13} = \nu_{23} = 0.25$ ), material II ( $E_1/E_2 = 25$ ,  $E_2 = E_3$ ,  $G_{12} = G_{13} = 0.5E_2$ ,  $G_{23} = 0.2E_2$ ,  $\nu_{12} = \nu_{13} = \nu_{23} = 0.25$ ). Except special mentions, for convenience, the following nondimensional parameters are used in numerical examples:

$$\begin{aligned}
\bar{w} &= \frac{100wE_2bh^3}{qL^4}, \bar{\sigma}_x = \frac{bh^2}{qL^2} \sigma_x \left( \frac{L}{2}, \frac{h}{2} \right) \\
\bar{\sigma}_z &= \frac{b}{q} \sigma_z \left( \frac{L}{2}, \frac{h}{2} \right), \bar{\sigma}_{xz} = \frac{bh}{qL} \sigma_{xz}(0, 0)
\end{aligned} \quad (11)$$

In order to verify the convergence of solution field, Table I presents variations of non-dimensional mid-span transverse displacement with respect to the number of series in  $x$ -direction ( $R$ ) and  $z$ -direction ( $S$ ) of  $0^\circ/90^\circ/0^\circ$  symmetric laminated composite beams. The results are calculated with  $L/h = 5$ , Material I and  $E_1/E_2 = 40$  for S-S, C-F and C-C boundary conditions. It can be seen that the responses converge quickly in  $x$ -direction and number of series in this direction  $R = 10$  can be the point of convergence of the displacement for the boundary conditions, whereas the beam appears softer and converges with an increase of the number of series in  $z$ -direction. As an example for further verification,  $R = 10$  and  $S = 4$  will be used in the following computations.

Static responses of cross-ply laminated composite beams is investigated in Tables II and III. The nondimensional transverse displacements at  $x = L/2$  are calculated for  $0^\circ/90^\circ/0^\circ$  symmetric and  $0^\circ/90^\circ$  un-symmetric composite beams with different boundary conditions and span-to-thickness ratios

TABLE I  
CONVERGENCE STUDIES FOR NORMALIZED MID-SPAN TRANSVERSE DISPLACEMENT OF  $0^\circ/90^\circ/0^\circ$  LAMINATED COMPOSITE BEAMS ( $L/h = 5$ , MATERIAL I,  $E_1/E_2 = 40$ ).

BC	S	R					
		2	4	6	8	10	12
S-S	1	0.9050	0.8794	0.8821	0.8827	0.8825	0.8826
	2	1.2946	1.2836	1.2867	1.2873	1.2871	1.2872
	3	1.2852	1.2765	1.2790	1.2793	1.2791	1.2793
	4	1.4654	1.4568	1.4594	1.4596	1.4594	1.4595
	5	1.4644	1.4559	1.4586	1.4587	1.4585	1.4586
	6	1.4647	1.4555	1.4582	1.4584	1.4581	1.4582
	7	1.4648	1.4556	1.4582	1.4584	1.4582	1.4583
C-F	1	2.2812	2.4497	2.5823	2.5810	2.5908	2.6161
	2	3.5129	3.8151	3.9580	3.9567	3.9663	3.9729
	3	3.5036	3.8114	3.9529	3.9528	3.9626	3.9693
	4	3.9105	4.2195	4.3625	4.3616	4.3713	4.3778
	5	3.9102	4.2183	4.3618	4.3620	4.3715	4.3777
	6	3.9163	4.2293	4.3739	4.3740	4.3833	4.3897
	7	3.9166	4.2295	4.3740	4.3741	4.3834	4.3897
C-C	1	0.6891	0.7536	0.8154	0.8482	0.8468	0.8469
	2	0.7495	0.8344	0.8962	0.9292	0.9278	0.9278
	3	0.7403	0.8287	0.8896	0.9222	0.9207	0.9207
	4	0.8256	0.9306	0.9919	1.0251	1.0235	1.0238
	5	0.8251	0.9294	0.9913	1.0243	1.0227	1.0235
	6	0.8306	0.9351	0.9970	1.0308	1.0285	1.0292
	7	0.8309	0.9353	0.9972	1.0309	1.0287	1.0293

TABLE II  
NONDIMENSIONAL MID-SPAN DISPLACEMENTS OF  $0^\circ/90^\circ/0^\circ$  AND  $0^\circ/90^\circ$  LAMINATED COMPOSITE BEAMS (MATERIAL II)

BC	Theory	$0^\circ/90^\circ/0^\circ$			$0^\circ/90^\circ$		
		$L/h=5$	10	50	$L/h=5$	10	50
S-S	HSBT [22]	2.414	1.098	0.666	4.785	3.697	3.345
	HSBT [6]	2.412	1.096	0.666	4.777	3.688	3.336
	HSBT [7]	2.398	1.090	0.661	4.750	3.668	3.318
	Quasi-3D [22]	2.405	1.097	0.666	4.764	3.694	3.345
	Quasi-3D [17]	2.405	1.097	0.666	4.828	3.763	3.415
	Quasi-3D [21]	-	1.097	-	-	3.731	-
	Present	2.418	1.105	0.666	4.918	3.730	3.346
C-F	HSBT [22]	6.830	3.461	2.257	15.308	12.371	11.365
	HSBT [6]	6.824	3.455	2.251	15.279	12.343	11.337
	HSBT [7]	6.836	3.466	2.262	15.334	12.398	11.392
	Quasi-3D [22]	6.844	3.451	2.256	15.260	12.339	11.343
	Quasi-3D [21]	-	3.459	-	-	12.475	-
	Present	7.077	3.496	2.257	15.889	12.452	11.333
	HSBT [22]	1.538	0.532	0.147	1.924	1.007	0.680
C-C	HSBT [6]	1.537	0.532	0.147	1.922	1.005	0.679
	Quasi-3D [22]	1.543	0.532	0.147	1.916	1.005	0.679
	Quasi-3D [21]	-	0.532	-	-	1.010	-
	Present	1.629	0.540	0.147	2.150	1.041	0.677

TABLE III  
NONDIMENSIONAL STRESSES OF  $0^\circ/90^\circ/0^\circ$  AND  $0^\circ/90^\circ$  LAMINATED COMPOSITE BEAMS (S-S, MATERIAL II)

Stress	Theory	$0^\circ/90^\circ/0^\circ$			$0^\circ/90^\circ$		
		$L/h=5$	10	50	$L/h=5$	10	50
$\bar{\sigma}_x$	HSBT [22]	1.0669	0.8500	0.8705	0.2361	0.2342	0.2336
	HSBT [42]	1.0670	0.8503	-	0.2361	0.2342	-
	HSBT [17]	1.0669	0.8500	0.7805	0.2362	0.2343	0.2336
	Quasi-3D [22]	1.0732	0.8504	0.7806	0.2380	0.2346	0.2336
	Quasi-3D [17]	1.0732	0.8506	0.7806	0.2276	0.2246	0.2236
	Quasi-3D [21]	-	0.8501	-	-	0.2227	-
	Present	1.1820	0.8668	0.7796	0.2564	0.2392	0.2335
$\bar{\sigma}_{xz}$	HSBT [22]	0.4057	0.4311	0.4523	0.9205	0.9565	0.9878
	HSBT [42]	0.4057	0.4311	-	0.9187	0.9484	-
	HSBT [17]	0.4057	0.4311	0.4514	0.9211	0.9572	0.9860
	Quasi-3D [22]	0.4013	0.4286	0.4521	0.9052	0.9476	0.9869
	Quasi-3D [17]	0.4013	0.4289	0.4509	0.9038	0.9469	0.9814
	Quasi-3D [21]	-	-	-	-	0.9503	-
	Present	0.4182	0.4613	0.4946	0.8068	0.8558	0.8869
$\bar{\sigma}_z$	Quasi-3D [22]	0.1833	0.1787	0.1804	0.2966	0.2911	0.3046
	Quasi-3D [17]	0.1833	0.1803	0.1804	0.2988	0.2982	0.2983
	Present	0.1262	0.1117	0.0880	0.0550	0.0683	0.0122

TABLE IV  
NONDIMENSIONAL MID-SPAN DISPLACEMENTS OF  $0^\circ/\theta/0^\circ$  LAMINATED COMPOSITE BEAMS (MATERIAL II)

BC	Theory	Fiber angle $\theta$						
		0°	15°	30°	45°	60°	75°	90°
S-S	Quasi-3D [22]	1.7930	1.8626	2.0140	2.1762	2.3030	2.3796	2.4049
	Quasi-3D [43]	1.7930	1.8626	2.0140	2.1762	2.3030	2.3796	2.4049
	Present	1.7933	1.8622	2.0119	2.1767	2.3091	2.3908	2.4180
C-F	Quasi-3D [22]	5.2683	5.4840	5.8705	6.2780	6.5930	6.7820	6.8442
	Quasi-3D [43]	5.2774	5.4898	5.8804	6.2879	6.6029	6.7919	6.8541
	Present	5.4658	5.6682	6.0602	6.4790	6.8093	7.0103	7.0769
C-C	Quasi-3D [22]	1.0866	1.1485	1.2616	1.3801	1.4711	1.5253	1.5431
	Quasi-3D [43]	1.0998	1.1537	1.2670	1.3856	1.4766	1.5309	1.5487
	Present	1.1842	1.2325	1.3446	1.4638	1.5559	1.6111	1.6293

TABLE V  
NONDIMENSIONAL MID-SPAN DISPLACEMENTS OF  $0^\circ/\theta/0^\circ$  LAMINATED COMPOSITE BEAMS (MATERIAL II)

BC	Theory	Fiber angle $\theta$						
		0°	15°	30°	45°	60°	75°	90°
S-S	Quasi-3D [22]	0.6370	0.6554	0.6608	0.6634	0.6650	0.6658	0.6661
	Quasi-3D [43]	0.6370	0.6554	0.6608	0.6634	0.6650	0.6658	0.6661
	Present	0.6369	0.6554	0.6608	0.6635	0.6652	0.6662	0.6665
C-F	Quasi-3D [22]	2.1599	2.2225	2.2402	2.2480	2.2529	2.2554	2.2562
	Quasi-3D [43]	2.1602	2.2228	2.2405	2.2483	2.2531	2.2557	2.2565
	Present	2.1593	2.2218	2.2396	2.2477	2.2528	2.2557	2.2566
C-C	Quasi-3D [22]	0.1367	0.1408	0.1431	0.1449	0.1462	0.1470	0.1473
	Quasi-3D [43]	0.1367	0.1408	0.1431	0.1449	0.1462	0.1470	0.1472
	Present	0.1362	0.1403	0.1425	0.1444	0.1458	0.1466	0.1469

$L/h=5$  and 50. The results are examined with earlier those derived from the HSBTs (Nguyen et al. [22], Khdeir and Reddy [6], Murthy et al. [7]), Quasi-3Ds (Nguyen et al. [22], Mantari and Canales [21], Zenkour [17]). It can be seen that there are differences of the transverse displacements between the present model and those from the HSBTs and Quasi-3Ds for the thickness beam ( $L/h = 5$ ), however the theories are similar with an increase of the span-to-thickness ratio. It can be explained by the fact that with an increase of  $L/h$ , the transverse deformation effects become smaller and the theories converge to the conventional ones. Moreover, it is interesting to observe that the present beam model clearly predicts lower stiffness than the previous ones for symmetric and un-symmetric laminated composite beams under all boundary conditions. Moreover, the nondimensional axial, transverse shear and transverse normal stresses of symmetric and unsymmetric cross-ply composite beams with S-S boundary condition are reported in Table III and compared with other works based on the HSBTs and Quasi-3Ds. In comparison, the effect of transverse normal and transverse strains are again found for the normalized axial stress  $\bar{\sigma}_x$ . For the transverse normal stress  $\bar{\sigma}_z$  and transverse shear stress  $\bar{\sigma}_{xz}$ , there exist deviations of the present theory and other one, especially for transverse normal stress. However, as expected the present solution tends to approach to the free-traction condition on the top surface of the beams.

As a mean to study effects of transverse normal and shear strains on the displacement and stresses further, Tables IV and V introduce variations of the nondimensional center transverse displacement with respect to the boundary conditions, span-to-thickness ratio and different fiber angles of  $0^\circ/\theta/0^\circ$  and

$0^\circ/\theta^\circ$  laminated composite beams. The results are compared to those obtained from Quasi-3Ds of Nguyen et al. [22] and Vo et al. [43]. Considerable differences between the present model and Quasi-3D ones are again observed for thick laminated composite beams ( $L/h = 5$ ) and no significant deviations are found between the theories for thin beams ( $L/h = 50$ ). Moreover, the increase of fiber angles makes the beam softer and leads to the increase of the transverse displacement.

#### IV. CONCLUSIONS

The authors proposed a new two-unknown model for static analysis of laminated composite beams. The axial and transverse displacements of the beam are expanded in a hybrid form under polynomial and trigonometric series. Lagrange's equations are used to derive characteristic equations of the beams. Numerical results for laminated composite beams with different boundary conditions are compared with previous studies and to investigate the effects of length-to-depth ratio, fibre angles and material anisotropy on the deflection and stresses of laminated composite beams. The obtained results showed that the proposed beam model is found to simple and efficient in predicting bending responses of laminated composite beams with various boundary conditions.

#### REFERENCES

- [1] M. KOMIJANI, J. N. REDDY, M. R. ESLAMI, M. BATENI, An analytical approach for thermal stability analysis of two-layer timoshenko beams, *International Journal of Structural Stability and Dynamics* 13 (08) (2013) 1350036.
- [2] K. Chandrashekhara, K. Krishnamurthy, S. Roy, Free vibration of composite beams including rotary inertia and shear deformation, *Composite Structures* 14 (4) (1990) 269 – 279.
- [3] C. Santiuste, S. Sanchez-Sez, E. Barbero, Dynamic analysis of bending-torsion coupled composite beams using the flexibility influence function method, *International Journal of Mechanical Sciences* 50 (12) (2008) 1611 – 1618.
- [4] T.-K. Nguyen, T. P. Vo, H.-T. Thai, Static and free vibration of axially loaded functionally graded beams based on the first-order shear deformation theory, *Composites Part B: Engineering* 55 (0) (2013) 147 – 157.
- [5] T.-K. Nguyen, B.-D. Nguyen, T. P. Vo, H.-T. Thai, A novel unified model for laminated composite beams, *Composite Structures* 238 (2020) 111943.
- [6] A. Khdeir, J. Reddy, An exact solution for the bending of thin and thick cross-ply laminated beams, *Composite Structures* 37 (2) (1997) 195 – 203.
- [7] M. Murthy, D. R. Mahapatra, K. Badarinarayana, S. Gopalakrishnan, A refined higher order finite element for asymmetric composite beams, *Composite Structures* 67 (1) (2005) 27 – 35.
- [8] M. Aydogdu, Buckling analysis of cross-ply laminated beams with general boundary conditions by ritz method, *Composites Science and Technology* 66 (10) (2006) 1248 – 1255.
- [9] M. Aydogdu, Free vibration analysis of angle-ply laminated beams with general boundary conditions, *Journal of Reinforced Plastics and Composites* 25 (15) (2006) 1571–1583.
- [10] L. Jun, L. Xiaobin, H. Hongxing, Free vibration analysis of third-order shear deformable composite beams using dynamic stiffness method, *Archive of Applied Mechanics* 79 (12) (2009) 1083–1098.
- [11] M. Simsek, Fundamental frequency analysis of functionally graded beams by using different higher-order beam theories, *Nuclear Engineering and Design* 240 (4) (2010) 697 – 705.
- [12] L. Jun, H. Hongxing, Free vibration analyses of axially loaded laminated composite beams based on higher-order shear deformation theory, *Meccanica* 46 (6) (2011) 1299–1317.
- [13] J. Li, Z. Wu, X. Kong, X. Li, W. Wu, Comparison of various shear deformation theories for free vibration of laminated composite beams with general lay-ups, *Composite Structures* 108 (2014) 767 – 778.

- [14] T. Vo, H.-T. Thai, T.-K. Nguyen, F. Inam, Static and vibration analysis of functionally graded beams using refined shear deformation theory, *Meccanica* 49 (1) (2014) 155–168.
- [15] T.-K. Nguyen, N.-D. Nguyen, T. P. Vo, H.-T. Thai, Trigonometric-series solution for analysis of laminated composite beams, *Composite Structures* 160 (2017) 142 – 151.
- [16] H. MATSUNAGA, Vibration and buckling of multilayered composite beams according to higher order deformation theories, *Journal of Sound and Vibration* 246 (1) (2001) 47 – 62.
- [17] A. M. Zenkour, Transverse shear and normal deformation theory for bending analysis of laminated and sandwich elastic beams, *Mechanics of Composite Materials and Structures* 6 (3) (1999) 267–283.
- [18] W. Chen, C. Lv, Z. Bian, Free vibration analysis of generally laminated beams via state-space-based differential quadrature, *Composite Structures* 63 (34) (2004) 417 – 425.
- [19] T. P. Vo, H.-T. Thai, T.-K. Nguyen, F. Inam, J. Lee, A quasi-3d theory for vibration and buckling of functionally graded sandwich beams, *Composite Structures* 119 (0) (2015) 1 – 12.
- [20] J. Mantari, F. Canales, Free vibration and buckling of laminated beams via hybrid ritz solution for various penalized boundary conditions, *Composite Structures* 152 (2016) 306 – 315.
- [21] J. Mantari, F. Canales, Finite element formulation of laminated beams with capability to model the thickness expansion, *Composites Part B: Engineering* 101 (2016) 107 – 115.
- [22] N.-D. Nguyen, T.-K. Nguyen, T. P. Vo, H.-T. Thai, Ritz-based analytical solutions for bending, buckling and vibration behavior of laminated composite beams, *International Journal of Structural Stability and Dynamics* 18 (11) (2018) 1850130.
- [23] S. Ambartsumian, On the theory of bending of anisotropic plates and shallow shells, *Journal of Applied Mathematics and Mechanics* 24 (2) (1960) 500 – 514.
- [24] E. Reissner, On transverse bending of plates, including the effect of transverse shear deformation, *International Journal of Solids and Structures* 11 (5) (1975) 569 – 573.
- [25] M. Levinson, An accurate, simple theory of the statics and dynamics of elastic plates, *Mechanics Research Communications* 7 (6) (1980) 343 – 350.
- [26] J. N. Reddy, A simple higher-order theory for laminated composite plates, *Journal of Applied Mechanics* 51 (1984) 745–752.
- [27] T. N. Nguyen, C. H. Thai, H. Nguyen-Xuan, On the general framework of high order shear deformation theories for laminated composite plate structures: A novel unified approach, *International Journal of Mechanical Sciences* 110 (2016) 242 – 255.
- [28] M. Murthy, An improved transverse shear deformation theory for laminated anisotropic plates, *National Aeronautics and Space Administration*, Washington DC, 1981.
- [29] M. STEIN, Nonlinear theory for plates and shells including the effects of transverse shearing, *AIAA Journal* 24 (9) (1986) 1537–1544.
- [30] M. Touratier, An efficient standard plate theory, *International Journal of Engineering Science* 29 (8) (1991) 901 – 916.
- [31] H. Arya, R. Shimpi, N. Naik, A zigzag model for laminated composite beams, *Composite Structures* 56 (1) (2002) 21 – 24.
- [32] C. H. Thai, A. Ferreira, S. Bordas, T. Rabczuk, H. Nguyen-Xuan, Isogeometric analysis of laminated composite and sandwich plates using a new inverse trigonometric shear deformation theory, *European Journal of Mechanics - A/Solids* 43 (2014) 89 – 108.
- [33] J. Mantari, A. Oktem, C. G. Soares, A new trigonometric shear deformation theory for isotropic, laminated composite and sandwich plates, *International Journal of Solids and Structures* 49 (1) (2012) 43 – 53.
- [34] V.-H. Nguyen, T.-K. Nguyen, H.-T. Thai, T. P. Vo, A new inverse trigonometric shear deformation theory for isotropic and functionally graded sandwich plates, *Composites Part B: Engineering* 66 (0) (2014) 233 – 246.
- [35] M. Karama, K. Afaq, S. Mistou, Mechanical behaviour of laminated composite beam by the new multi-layered laminated composite structures model with transverse shear stress continuity, *International Journal of Solids and Structures* 40 (6) (2003) 1525 – 1546.
- [36] M. Aydogdu, A new shear deformation theory for laminated composite plates, *Composite Structures* 89 (1) (2009) 94 – 101.
- [37] K. Soldatos, A transverse shear deformation theory for homogeneous monoclinic plates, *Acta Mechanica* 94 (3-4) (1992) 195–220.
- [38] N. E. Meiche, A. Tounsi, N. Ziane, I. Mechab, E. A. Adda.Bedia, A new hyperbolic shear deformation theory for buckling and vibration of functionally graded sandwich plate, *International Journal of Mechanical Sciences* 53 (4) (2011) 237 – 247.
- [39] S. S. Akavci, A. H. Tanrikulu, Buckling and free vibration analyses of laminated composite plates by using two new hyperbolic shear deformation theories, *Mechanics of Composite Materials* 44 (2) (2008) 145.
- [40] J. Mantari, A. Oktem, C. G. Soares, A new higher order shear deformation theory for sandwich and composite laminated plates, *Composites Part B: Engineering* 43 (3) (2012) 1489 – 1499.
- [41] C. H. Thai, S. Kulasegaram, L. V. Tran, H. Nguyen-Xuan, Generalized shear deformation theory for functionally graded isotropic and sandwich plates based on isogeometric approach, *Computers & Structures* 141 (2014) 94 – 112.
- [42] T. P. Vo, H.-T. Thai, Static behavior of composite beams using various refined shear deformation theories, *Composite Structures* 94 (8) (2012) 2513 – 2522.
- [43] T. P. Vo, H.-T. Thai, T.-K. Nguyen, D. Lanc, A. Karamanli, Flexural analysis of laminated composite and sandwich beams using a four-unknown shear and normal deformation theory, *Composite Structures* 176 (2017) 388 – 397.

# Water, Urban Morphology, Urbanization and Sustainable Development: Case study of the Nhieu Loc – Thi Nghe canal area in Ho Chi Minh city

Xuan Son Do  
Faculty of Civil Engineering  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
sondx@hcmute.edu.vn

**Abstract**—Water is the source of life, forming of the urban space, cultural values, customs, habits, and a catalyst for human civilization. Similar to the major cities of the world, Ho Chi Minh city, which has been proven as a city born from the water, has formed and developed based on water. But nowadays the symbol on water in defining the image of city is impacted by processes of urbanization. Notably, the urban surface water area is being encroached upon and gradually narrowed down due to the demand for economic development and infrastructure. This issue is related to the pollution of water surface, environmental degradation, flooding from increasing run-off and the rising temperatures in the urban areas. The article explores the water interface in specific peri-urban areas in the Nhieu Loc – Thi Nghe canal on both sides. The current situation of urban surface water use and water quality in this local reflect various changes in the Ho Chi Minh city's environment, lifestyle and urban morphology in general when it has completely gone beyond the geographic limit and come into the city. With the related factors of changes in trade activities, the landscape of urban social cultural architecture, etc. with a water interface, the article analyzes the causes and evaluates the problem-solving approaches from the urban projects for urban water surface improvement.

**Keywords:** Nhieu Loc–Thi Nghe (NL–TN) canal, Landscape, Urbanization, Pollution, Waterway, Urban morphology

## I. INTRODUCTION

### A. Saigon and the water

Society, space and urban landscape, the mirror of the river: The close link of human beings to the watercourse has finally well evolved over the centuries. « Rivers bear the mark of a society; they are its mirror. In other words, each stage of the society corresponds to one state of the river » [1].

Water was omnipresent in the town of Sai Gon which corresponds to what GARNIER says «the symbolic force of water in the urban image» [2]. It is surely still true for Saigon today, but probably much less than before. Saigon – Ho Chi Minh city (HCMC) is bordered by the Sai Gon River and by the Nhieu Loc – Thi Nghe (NL – TN) canal in the North, by the Ben Nghe canal in the South.

This “fragile” balance of the land, water and human give Sai Gon an original identity of the landscape, organically connected to human life. Man and water are inextricably linked together; the town of Saigon was effectively born from water, made of riverbanks and lakes. The Vietnamese past showed a rich symbiosis between the city and the river: we

speak about the « couple city – river» or of the “couple man – river”.



Fig. 1. Scheme of the TN-NL canal (Source: Author, 2019)

As for the characteristic landscape image of the area of Ha Noi as a comparative case study, HCMC is also a city of water of rivers, arroyos, marshes, lakes or ponds bustling with "quays and boats". Until today in the territory of HCMC, there are about 3,000 km of rivers and canals [3]. According to some ancient documents, the Thi Nghe Arroyo was a natural geographical boundary between the inner city and the suburbs of Gia Dinh. Similar to the geographical character of the To Lich river in Ha Noi, the NL – TN canal is the natural waterway crossing the central part of the city and flowing through to the following seven districts of 1, 3, 10, Phu Nhuan, Tan Binh, Binh Thanh and Go Vap with a total area of 36 km<sup>2</sup>, 8.7 km in length, average width of 27 m in the upstream, and 90 m in the downstream, the average depth of the canal is 5 m. Canal area is 12 km<sup>2</sup> wide [4] (Fig 1). It flows from West to East and its winding course borders the left side of the Old Gia Dinh Citadel before and of the downtown today.

Nhieu Loc is the section from Thi Nghe bridge to upstream (bounded by the Ut Tich road, district of Tan Binh), and the part from Thi Nghe to the Sai Gon river, near Bason port is called as the Thi Nghe arroyo. Along with the Sai Gon River and the Ben Nghe – Tau Hu canal, the NL – TN canal is one of the three most important natural river channels of Saigon – Gia Dinh till today (or these days).

### B. Evolution over time of the river, its symbolic representation and problems.

To understand the changes of the NL – TN canal, it is necessary to go over again the geography and the history as well as culture of Sai Gon through the place and the relations of the river with the city. At the beginning of a feudal regime, the Citadel was controlled by a king and mandarins; the town and the trade villages on the river were only one complex.

Another important dimension that should not be ignored from this watercourse is from its main values to shape a landscape of the river with its space. This new structure of the



river could help people find their childhood again, which has made this landscape as the bridge to link the past to the future. However, this landscape was space of leisure, relaxation and meditation but it also created a cultural and historical local spiritual space, a space of research and contemplation, etc.

Facing this reality, the researchers raised some questions about the acute problems and study the following questions: "What are the landscape values of the river that still subsist today? What are the important elements to preserve and how to highlight the landscape of the NL – TN canal while these elements were and are alive, still preserving city memories, beauty, landscape, and environment? It is therefore of great importance to give back to the river all of its values today by finding answers to the questions posed by the various components of the river-side urban landscape.

The problematic concerns more precisely the current landscape problems, the renewal of the relationship and the integration of watercourses in the urban network. It is therefore important to understand not only the impacts of urbanization on the river, but also, conversely, to understand the influence of the rivers on and their contribution to the surrounding urban spaces.

There have been many civil and practical projects, namely: built two traffic lanes, reinforced both sides of the river with concrete to protect the river space, improved water sources, strengthened the eco-system, built the landscape on both sides, etc. But there is currently no specific research project focusing on heritage, landscape environment or urban morphology of the NL – TN canal space.

### C. Objectives of study

The study of the urban river landscape aims to understand how the urbanization in HCMC has influenced the river landscape as well as the villages. From there, the researchers will look for the specific landscapes of the NL – TN canal. These are landscapes of former villages and current urban villages, which define the landscape values and characteristics of the NL – TN canal today. It is about the landscape identities that are highlighted, revitalized and integrated into the present life. Subsequently, it will be necessary to represent and implement this research on the plans and on the elevation of the urban landscape of the villages and river. This is in close association with the social activities on the two banks: the NL – TN canal morphology, the heritage and the traditional dwelling space with its historical and cultural values, and the urban composition of the NL – TN canal. It also links to the environment, the social space, the spaces of worship and recreation activities, prosperity of the city, commercial activities and local economic sectors in relation to river traffic, etc. in favor of the HCMC.

The study took into account the different dimensions of water, its importance in the fluvial ecological landscape of the city as well as the perception of this water by the population aims to cast a light on the omission of the importance of this water, and in the future, it is necessary to set up a duty and regular supervision of the need as well as to meet for a requalification of the NL – TN canal.

### D. Methodology

The author used methods of literature review, analyzing ancient images data and documents related to regional urban history and culture. In addition, the site survey is also applied

for analyzing and assessing the geographical status, landscape and urban society of the study.

The literature review method is applied for introducing and emphasizing the problematic of research, demonstrating the urban transformations.

The site survey is used for assessing the actual situation and diagnosing the values of water space in interaction with the city. The analytical method leads to the synchronization of the research results.

## II. LANDSCAPE REPRESENTATIONS WITH LOCAL CULTURE AND THE SHAPE OF VILLAGES NEAR THE WATER

Where did the habitations settle? It certainly exercised such a strong influence in this low-land part of the country, threatened by the floods that according to P. GOUROU's observations, people tended to gather on high-ground areas, which were only affected by heavy floods; we will see later that the relief reveals guidelines (river and affluent) in the dispersion of the Tonkinese villages. Still in almost every habitation, people often prefer to build religious buildings and their homes near water bodies, such as river, lake, pond or near water sources. To defend themselves against the external attacks or invasions, they had to rely solely on their own strength; hence the desire to group and surround the village with a solid bamboo fence [5].

These elements, which reflect some design conventions of construction associated with community activities, partly explain the landscape layout as a type of traditional village based on the river.

Many villages are subdivided into hamlets named *thôn* in Vietnamese language.

Regarding the form of the village territory: "The villages located in the major river beds and the coastal zone, for which the geometrical forms are applied: their territory is generally composed of a band perpendicular to the river or the shoreline. These villages of rather recent formation have either to touch the river, in the hope that a favorable displacement of its bed will ensure an extension of their territory, either touching the sea, in order to benefit from the formation of the foreshores"[6]. In regards to the form of the villages, "they form bands parallel to the river and sometimes continue for several kilometers, marrying all sinuosities of fluvial bed; the sharpness of their design allows them to be distinguished at first sight from the villages of major river bed, which have the forms much more confused"[7]. Houses have been built on the high ground, beside the narrow water space, and village streets along the water.

In our study related to the landscape of the NL – TN canal, to analyze the landscape on a territory and to highlight it, the first important thing to do is to understand concepts and definitions of the landscape, the heritage, the planning of the territorial landscape, etc. the attention must then be focused on understanding the culture, history, social space, life styles, living environments of the river and its neighboring populations constituting the urban landscape. This documentary research process provides the foundation for theories, concepts and knowledge. Also, it is later the method of analysis of the "pillars" on a territory. X. BROWAEYS and P. CHATELAIN express that "landscapes, heritage, planning and environment are therefore the four pillars of an analysis of the municipal territory. To carry out this analysis, fieldwork

is essential. But also to collect data, to use statistics and decrypt the map, document whose purpose is to restore the landscape on a precise scale" [8]. And this is useful for a city or a rural territory as well as for the natural component of the landscape to be investigated.

### III. ANALYSIS OF SITES LINKED TO THE RIVERS UNDER THE URBANIZATION IMPACT

The role of the TN-NL canal, "advantages and potentials" for the landscape of Saigon – Gia Dinh in the past and today. The history of the construction of Gia Dinh Citadel began more than 340 years ago when the palace Mayors of the Trinh Family (in the north) and the Palace Mayors of the Nguyen Family (in the south) engaged in the fight for power and territory, the enemies of the South then invaded the land of the Nguyen Mayors, at the time led by Nguyen Anh. The latter led the battles to the south, in the Kingdom of Cao Mien, and continued the combats until the enemies withdrew [9]. In 1790, he built the wall of Gia Dinh to occupy and keep his area. Then, the Minh Mang King had demolished the Citadel and rebuilt it in 1835, and named it Sai Gon. At the time of French colonization of Gia Dinh in 1861, the French had destroyed the old Citadel of Sai Gon in 1859.

The Citadel of Turtle (Gia Long built it in 1790 and Minh Mang destroyed it in 1835) was called the Gia Dinh Citadel, the composition of this Citadel was the "diagram of the eight divinatory signs" according to the mixed architectural types of East-West, drawn by the French man Olivier de Puymanel. The Citadel of Phoenix (Minh Mang built it in 1836 and the French destroyed it in 1859) was called Saigon Citadel; it is the smallest construction of the "diagram of the eight divinatory signs" of architectural type "Vauban" [10].

Thus, when we must distinguish the two Citadels, we refer to their names: Citadel of the Turtle (Gia Dinh) and Citadel of the Phoenix (Saigon). Indeed, when we talk about the Citadels that are built, we refer to the general names: Citadel, Saigon Citadel Gia Dinh, or Citadel Phiên An.

The observations of the oldest map of Gia Dinh Citadel in 1795 show: (Fig 2)

The NL – TN canal is located in the east and northeast of the Gia Dinh Citadel. With such a favorable position, the NL – TN canal with Tau Hu - Ben Nghe canal and Sai Gon river are responsible for acting as a natural wall of "water" surrounding the outer ring of the old Gia Dinh Citadel and were the green lungs and transportation for residents living in the Citadel.

#### A. The role of arroyos for the historical and cultural heritage:

The arroyo system had existed long before and had links with the life of the previous owners from the previous country. However, for the Gia Dinh Citadel, since 1674, during the widening of the borders towards the South to reach this region, the arroyo became closely related to the formation of the new region of Saigon - Gia Dinh from An Nam [11]. This means that it has existed for over 340 years since the reign of the Nguyen dynasty in the history of Vietnam.

It was a former territory of Cao Mien, which according to historical proof was one of the most prosperous and dynamic cities of Indochina. For a long time, a significant number of Chinese came to settle there for trading purpose and building houses as well as worship buildings (temple, pagoda),

maintaining the intellectual and spiritual life to remain there as long as possible. They shared with the Vietnamese and part of the Khmers, the living space on both banks of the canal. As a result, the cultural history of the NL – TN region or as Ben Nghe – Tau Hu canal exemplifies the mixing of identity of each from the three cultures.

Another testimony about the NL – TN canal, it is possible to observe three types of historical worship buildings with different architectural styles of three ethnic groups on the same arroyo: Khuong Viet Pagoda (the Buddhist school of Mahayana) with the Vietnamese culture and architecture, Nam Tong Khmer Candaransi Buddhist Pagoda (the Buddhist School of Hinayana) with the Khmer architecture, Vinh Nghiem Pagoda with the Vietnamese culture mixed with that of the Chinese and Phuoc Hai Pagoda (Ngoc Hoang Palace) with cultural traits Chinese mixed with Vietnamese culture, etc.

#### B. The geographical role, the strategic meanings in terms of military defense and historical values of the river:

Like rivers of Thang Long Citadel, rivers, lakes and ponds embrace and protect the Citadel. At the time of Gia Dinh Citadel construction, our ancestors made use of the natural geographical features, such as its strategic location, surrounded with the complex system of canals and marshes playing the role of both defense and attack to protect the Citadel:

- In the North East: the NL – TN canal undulates at the foot of the walls (the map in 1795 clearly shows the old wall leading close to this river),

- In the South West: the Ben Nghe - Tau Hu canal (commercial district of the Chinese and Japanese),

- In the South East: is the Saigon river (otherwise called the Dong Nai river),

- In the North West: is the TN-NL canal which flows behind the large area of marshes and agricultural sector of food plantations for the Citadel.

The arroyo-and-canal network is like a natural water wall, the second military outer ring area that flows along the bottom of the Citadel and protects it.

According to the map of the Gia Dinh Citadel dated back to 1795, the transportation system and important routes of advancing and retreating for military defense are based on water lines (guard and attack positions). Road traffic is not developed yet at this time, the entrances and exits through the gates to the Citadel were only passed by means of these waterways.

Since the Emperor Nguyen Anh built the Citadel of Gia Dinh, Thi Nghe canal has become an important place of this territory, because with the Saigon river, these two rivers form a complementary angle protecting Gia Dinh Citadel. The water of the surrounding ditches was led from the arroyo and the beginning of that arroyo is chosen to build a workshop of warship repairs of the Nguyen naval army, called the Chu Su (or Ba Son) workshop [12].

#### C. The directing role of the canal and arroyo in urban planning and composition:

At the time of construction of Gia Dinh Citadel by Gia Long or that of Sai Gon by Minh Mang King, the builders used the configuration of a Vauban-style square influenced by the

Western ideology of walls planning (because at the time there had been so many back and forth of Western traders who were in close economic and political relationship with the Vietnamese Court). Inside the Citadel, buildings were organized according to the circulation of a chessboard with the image of the inner Citadel. At its exterior, the spatial composition of the old city is prescribed by taking the Citadel as its centre and in drawing the chessboard to go in all directions. To do this, the main canal and arroyo line have been used as the essential directional axis of surrounding development and widening ways of the city in different directions.

The main river Dong Nai – Sai Gon plays a decisive role in both the main and secondary orientations for the Citadel of Gia Dinh in terms of geomantic orientations for a sustainable development. In the same way, the branching arroyos of the NL – TN canal located on the left and that of Ben Nghe – Tau Hu on the right were part of spatial and directional composition of Gia Dinh – Saigon Citadel.

#### *D. The directing role of the canal & arroyo on professional cultures:*

The To Lich – Kim Nguu rivers and the Red river surround Thang Long Citadel in the North. Nearly a hundred years ago, the NL – TN water ways, Ben Nghe and Tau Hu wharves, as well as the Dong Nai river formed the natural boundary dividing the inner city of Saigon – Gia Dinh one side and one area of the Citadel on the other, in terms of urban and professional life.

The inhabitants of the inner city are mainly engaged in trade and service activities, traditional handicrafts of high quality and ingenuity. Outdoors, it is mainly farmers carrying out the work of cultivation, breeding of aquatic products, seafood as well as some craft trades.

The element of land and water of the southern delta reflects certain differences from the northern region. Firstly, the geographical character of the southern fluvial region is the low marsh band. The habitable solid high zone is very far from the river. Secondly, the geological, geographical and topographical characters with sedentary habitat culture are extremely simple among farmers in the South; it is enough for them to drive three stakes to temporarily set up a straw hut where the family lives in a mobile boat in complete freedom without having a solid house on any land plot to remain sedentary for a long time.

Thus, all along the canal, it may not be able to see many settlements of poor farmers living on agriculture and fishing activities - if not very precarious on the land. It is only possible to see their boats covered with wicker wandering here and there, which constitute both their shelter and their means of transport. The zone of sedentary habitations in the villages is quite far from the river (different from the sedentary culture of Northern residents who appreciate the proximity of the waterline).

#### *E. The directing role of the canal on economy:*

The Saigon – Gia Dinh Citadel was an economic center of the country from the time of Gia Long King because it enjoyed a very advantageous geographic location near the river and waterlines, main transportation routes at the time. From inside of the provinces, boats could follow smaller or larger rivers to enter Gia Dinh – Saigon along the Mekong river to reach Dong Nai, at Ben Nghe estuary and the Nhieu Loc river to enter Gia

Dinh. The big boats of western traders (Portuguese, Spanish, Dutch, English, French, etc.) coming to Asia (to China, Japan, etc.) could follow the maritime way, entered the mainland through the river estuary of Dong Nai - Saigon, arrived at Gia Dinh Citadel and stopped at estuary of Dong Nai river. Also, since the time of Gia Long, there were commercial streets in the region, such as the Chinese street and the Japanese street on the banks of the Ben Nghe – Tau Hu canal of the Cho Lon (Grand Market); they are still recognized today with shops and services. As a result, a familiar image has been shaped, which has become the trademark of the fluvial and commercial landscape city: "Ports and boats". A tumultuous animation praised the numerous commercial activities both on the boats (in the canal) and on its banks. (The perspective drawing of Sai Gon Citadel old painting by M. A. Lepère made in 1881 showed part of the dense city of boats buying and selling on the river) (Fig 4).

The territory of Saigon is located towards the end crossing grand rivers sources or arroyos, a fertile area covered with alluvium coming from the upstream region. The climate is mild throughout the year with dry and rainy seasons, favorable to farming activities. Thus, the agricultural economy here has all elements required for the land development for agriculture, farming and fishing.

#### *F. The social role of the canal and the arroyo:*

In Saigon, the Ben Nghe arroyo, the Tau Hu canal and the NL – TN canal, create the border between the city and the countryside, the gateway and exit of the capital and the connection with adjacent provincial areas. At the time when the motorized vehicles were not yet developed, the getting around and transport were realized essentially by means of the waterways, this arroyo formed part of everyday life of the Saigoners. It has been the meeting point, social contact, trade and merchandize with the traders coming from all over the region. The place was also appropriate for meetings, social contact establishment, enrichment of communication and opening to the outside world. It was a space for exchange, learning, connection, social cohesion between the land and the canal.

The waterspace is also a place of celebration not only social activities related to trade exchange, meetings, but also traditional cultural festivals, worship activities, both inside and outside villages: floating lanterns ceremony, the Buddha birthday's celebration, the pogodas of Vinh Nghiem, Hai Duc and Van Tho. It was a space of cultural and spiritual activities within social activities.

To clearly define the geographical location and Feng-shui axis of Gia Dinh Citadel that our ancestors made, it is necessary to provide arguments recognizing the Feng-shui geography of Sai Gon Citadel. Regarding the geographical localization and the landscape area surrounding Gia Dinh Citadel, elements would be inspired by the Feng-shui scientific logic, especially when we examine the ideal model of the School of Forms in the Feng-shui [13].

It has been confirmed that Gia Dinh or Saigon Citadel - is built in a good land location, a land that our ancestors had well chosen to develop the city of Gia Dinh. Thus, the "Thanh Long" - Blue Dragon - is the NL – TN canal undulating to the left of Gia Dinh Citadel, then continuing its course like a big arm "Thanh Long" always on the left of the river. The hills (green area on the left side of the map) serve as support while accompanying the "Thanh Long". The "Bach Ho" - White

Tiger - as the Ben Nghe - Tau Hu canal, is located on the right of the good land with nice view point; he pushes it, embraces it and bends towards the Gia Dinh Citadel (show on the Fig 3 of the Feng-shui composition).

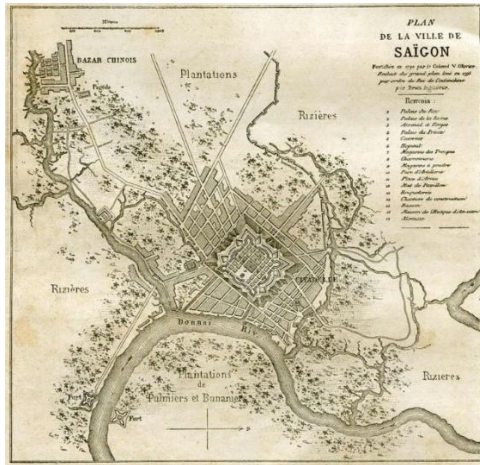


Fig. 2. Old map of Gia Dinh Citadel built by Gia Long with the NL – TN canal, oldest map of the location, drawn in 1795.[14]

+ The directing role of the river on the geographic composition and on spiritual Feng Shui: as mentioned above and according to the maps of urban planning and transport before and today, the city develops in the form of a chessboard with a parallel widening orientation from Gia Dinh city. When determining the geographical and geomantic orientation of Saigon, the ancestors chose the location of the old Citadel of Saigon – Gia Dinh as the center to define the way in which the directions were linked to the spiritual axis of Feng-shui. (Fig 3)

The axis of geographical orientation of Feng-shui was the orientation of the Gia Dinh according to the urban transport development axis towards the Dong Nai river and Saigon estuary.

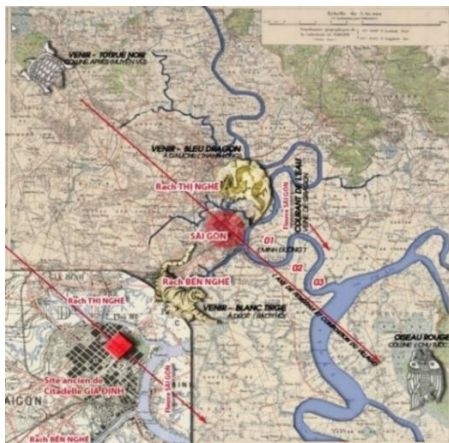


Fig. 3. Feng-shui structure of Gia Dinh Citadel and Saigon city's Diagram in the past with support of Saigon surrounding area's map in 1911(Source: Author, 2019)

On both sides, surrounded and protected by two mountain ranges; on the left, from the Blue Dragon "Thanh Long" to the NL – TN canal, on the right, from the White Tiger "Bach Ho" to the Ben Nghe - Tau Hu canal;

In this case, it is possible to say that the NL – TN canal is a main water element of the Feng-shui of Gia Dinh or Saigon from the construction of the Citadel until today.

#### G. The role of the canal in the natural landscape and the ecological environment of the city:

The NL – TN canal was a "green lung" of the Gia Dinh - Saigon Citadel. The drawing of Saigon in 1881 proposed by Mr. A. Lepère clearly shows this (Fig 4). It has been an ecological landscape axis of the vast fluvial region next to the urban city, joining with the immense outer space of the ecological zone of villages and countryside in the suburbs of Saigon city. Saigon in the old days was described as a bustling city of wharves and boats crossing and omnipresent going and returning over the entire city. Since the middle of the twentieth century, the NL – TN canal has gradually lost its circulation function and retains only that of drainage for the agricultural zone of neighborhoods around the city. The ecological landscape of the Nhieu Loc canal, with its limpid, green and fresh water, arouses a feeling of nostalgia among many Saigon people who became old.



Fig. 4. Drawing of Saigon in 1881. Saigon is a city with water landscape, "floating city". According to Nature /by M. Favre; Navy Infantry Captain... [15]

After the Second World War, people from everywhere rushed to the city to look for job opportunities and new lives. As the city was more and more overcrowded, the ecological open space "green-blue" (the floating city) continued to be restricted. From the years 1960 to 1990, due to the urbanization, the density of construction along the river systems increased and occupied the ecological spaces on the banks of arroyo and rivers. The ecological natural landscape space was reducing and its pollution - more and more serious - was gradually impacting human health and causing social instability. It was time for the local authorities and inhabitants to think about bringing life back to the river, to restore some of these spaces to this river and its residents. The project to upgrade and renovate the landscape of the NL – TN canal began, and in August 2012 phase 1 of restoration and renovation of the river landscape finished. The author will come back to this issue in the following section 4.2 on the concrete analyses of the evolution process of the NL – TN canal zone as well as the impacts on the landscape and urban ecology.

#### IV. RECOMMENDATION FOR THE URBAN LANDSCAPE OF THE TN-NL CANAL

According to our surveys and site studies and based on the study of documents concerning the subject, it is important to notice that the renovation of the NL – TN canal allowed to improve the landscape, to expand its stream, to revitalize the river and to reintegrate it into the society (among the Saigon inhabitants). This planning permitted also an enhancement of the city image with a landscape and architecture in a better



harmony with the environment. However, only the first phase of this process has been completed. The second phase should be continued to separate the river stream from the sewage collection network so that the water will be treated before being discharged into the Saigon river. The aim is to make its water as a source of urban ecology, cleaner, and thus allow the inhabitants to access it directly through the cultural activities and the leisures.

In this direction, the author would suggest below the general diagram of the main landscapes entities of the NL – TN canal which should be emphasized and promoted:

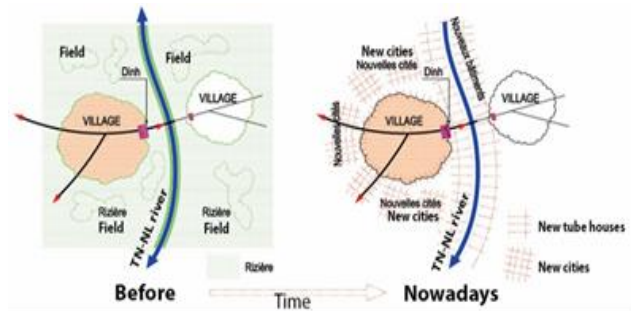


Fig. 5. Landscape morphological change of villages and canal (Source: Author, 2019)

#### A. The existing landscape entities to be promoted

As a result of the analysis and assessment process, it is time to affirm the actual significant values of the NL – TN canal representing on the landscape of the fluvial street in terms of urban morphology in its role of a peripheral ecological space of HCMC. And furthermore, this watercourse is a conjunctive and zoning artery of functional areas, separating the city center from its suburbs.

Regarding the traffic function, it is important to notice that the roads network crossing place of the city's major thoroughfares and bridges with the northern, northeastern and current central areas, occupies a strategic location in the construction of the urban image: beauty of the city on the stream edges, uses of water element itself.

This river-side street is also the place where the spirit and the memories of old villages and traditional crafts are maintained. Therefore, it is necessary to reinforce the social relations between the communities of old fluvial villages now becoming urban residential areas. It is possible to re-inspire from the cultural identity in the past of this place with animated activities to renew the social connection along this fluvial urban axis.



Fig.6 . Today, on some old sections of the Thi Nghe arroyo, there are still some precarious settlements illegally raised during the years 1960-1990. It is possible to keep a part of it to trace a historical aspect of the landscape and thus make it an identity of the sector. (Source: Author, 2019)



Fig. 7. Make the buildings more expressive, take into account the bridges, important supporting element of the landscape, integrate them into each space to bring an identity to the landscape and an expressive color to the waterway. (Source: Author, 2019)

#### B. The forgotten landscape elements need to be improved and promoted

##### 1) The public spaces

##### a) Public spaces in the river-side villages

Conventionally, the spaces of the first floor of the individual house overlooking the street are used for shops, business offices, a place for commerce or for retail. The circulation axis and the shops row obviously become commercial place - space of communication with the public area. The larger the circulation axis is, the more important and animated the public space will become.

This public space is noted for cultural characters, different purposes of use and professional activities, the characters of this space change throughout the day as graphically illustrated below:

As being located nearby the river with high traffic density, all the buildings along the road proposed many types of commercial services such as restaurants, hotels, etc. They kept open early in the morning until lately in the evening in responding to the public's demand. The public spaces are therefore more diversified: commerce, meetings, traffic, exercise, entertainment, relaxation and spiritual and festive activities, etc. These public spaces should be highlighted in respect of the river landscape and urban ecology so that they should become an attractive place for Saigon inhabitants and for the city tourism.

##### b) Improvement of public spaces along the canal (Fig 8)

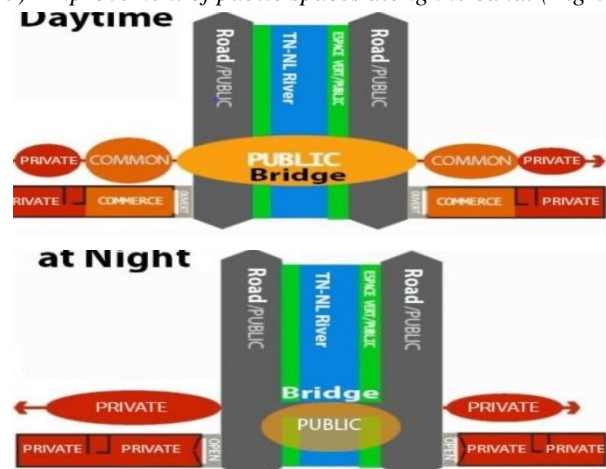


Fig. 8. Public spaces on Truong Sa street and Hoang Sa street along the TN-NL canal (Source: Author, 2019)



The public space is inside the "private" space: this mixture of use of space between public space and private area is often observed with the buildings located along the river street. As a consequence, the businesses grow on a large scale, until occupying the main public spaces with many conflicts. This way of use of space spontaneously developed in families by the lack of planning management should be better controlled by the competent authorities.

The other public spaces along the river banks such as sidewalk space, pedestrian crossing way, garden and vegetable lawn, etc. even planned, are illegally occupied by motorized vehicles, shops, restaurants, etc. They are usually observed to be impacted by the particular interest so seen in the middle of dust, noisy horns, pollution, etc. with messy concrete construction and architecture, less greenery, etc. due to the lack of a synchronous and strategic urban planning. It is important to improve the riverside urban plan that can promote the public space's values and to sanction any type of encroachment of public space in order to revitalize his inherent social role for the city and the citizen.

## 2) The commercial activities

The river-side business was formerly called "chợ bến" (market on the quay) by Saigon inhabitants with the image of the landing and leaving boats and economic, social, cultural and political space. Inside the river-side villages, the chợ bến was found at the starting point of each village which communicated with the waterline and connected to the other sections of the city.

Today there is a fundamental change of the central section of the river. The environment is impacted by polluted water, trash of modern life, traffic, etc. The markets were completely transformed. Although they remained close to the river and public space, but people do not use any more the river for the transportation of goods and circulation.

On the further section of the NL – TN canal, some rural markets still exist near by the river station or old village public space, the commercial activities are facilitated by peasant's own rural characteristics: street vendors with the and two baskets, goods displayed on the ground, etc. (Fig 5 & Fig 8).

The role of "chợ bến" or markets is currently observed as totally confused in the middle of the contemporary urban complex while they deserve to be improved and highlighted. Thereby, it is crucial to restore the inherent circulation and transportation function of the NL – TN fluvial axis which will connect these markets together, so let the Saigoners access directly to the water and thus connect the local citizen with the outside. It is possible to think of developing the local economy through fluvial transportation connecting cultural heritage and landscape, park, restaurant, hotel, etc. with the waterway eco-tourism and for the future urban development.

Other elements suggested to be improved for the river's landscape

For temporary shops and street vendors:

- It is recommended to respect the regulations of urban landscape design by avoiding the temporary kiosks and controlling the urban signage with specified size and color in order to improve the city image.

- The street vendors should be allowed to operate only in certain schedule and location: market, park, station with a strict urban hygiene regulation. The schedule should be

limited during the peak hours with dense traffic to avoid the traffic jam and the concentration of people in the public space.

- It should have a better control of all small shops along the river as well as on the circulation axis that discharge in the living environment a quantity of wastewater, organic and inorganic trash without any recycling or treatment process. Fighting against the pollution from the various waste accumulated in the river is something to do to improve the quality of the NL – TN water and the air of this area.

For the sedentary stores:

- These shops are often located on the three first floors of the tube house or the office building. They are not the same as there are any common norm applied to the urban architecture and the facade management: various front views, different canopies, different sizes and colors of signposts, etc. One strict control with a clear guide book and a sanction chart should be applied and followed up by the competent authorities of each section.

- It is necessary to forbid the grocery shops to occupy the sidewalks for their private business purpose. The vehicles should be parked correctly separating with the appropriated walkway for the pedestrian to contribute to solving the problem of traffic jams, circulation and to improve the visual and functional image of the fluvial axis.

## C. Proposal of the urban design model: Preservation and restoration of a characteristic landscape of the canal (Fig 9)

Identify the values by improving them, target the modalities of the landscape development of the river-side spaces, that on the waterway include three pillars: Society, Economy and Environment

According to the analysis carried out, the landscape structure of the village in the direction of a sustainable development consists of three main elements: the Society: Culture – History - Heritage and the Economy- Traditional Trades and Crafts – Agriculture, the Ecology: Sustainable Life - Habitation - Atmosphere, elements of identity of the characteristic landscapes. (Fig 9)



Fig. 9. Identification of the characteristic landscape structure of a village on the canal as well as the canal. (Source: Author, 2019)

The circulation network represents both the connecting line between the separated elements to compose an almost complete landscape spatial plan, and the essential link with the inhabitants to favor their access to this space of the canal.

The three groups of elements are found in the core of the villages, each of which has different contents, but their structure is considered analogous to these three groups of main elements which bring their effects in different ways to the landscape (proportion, scale, frequency of urbanization, etc. subjective - objective and direct - indirect interventions of man).

Our research made it possible to show the problems of the current urbanization in Saigon which brings their impacts on the landscape structure of the NL – TN canal, impacts on the characteristic values of the traditional villages which created a landscape structure elevation bordering the NL – TN canal.

## V. CONCLUSIONS

With regard to this important content, our studies and analyzes have led to answer to the question: "What are the urban landscape values of the NL – TN canal in the past and today? And how to promote them into the current city's life?"

We focused our studies on two main dimensions:

+ Dimension related to urban sociology from the point of view of geography and landscape of the population.

+ Urbanistic dimension of the morphology, structure and composition, cultural and commercial activities of the current NL - TN canal.

For the NL – TN canal fabric and material study approach, we consulted and used urban analysis methods by inspiring concrete works such as K. LYNCH's "Image of the city [16]," "Elements of Urban Analysis and Urban Forms, Urban Fabrics by P. PANERAI [17], "Morphology, Geography, Planning and Architecture of the City" by Alain ROGER [18], "Spatial Analysis, Cartography and Urban History" by J. Luc ARNAUD [19], to analyze the landscapes of the NL - TN canal.

Our different fields of analysis are related to the urban landscape, the architecture, the urban design, etc. gathered in a territorial landscape of NL – TN canal.

The research allowed the readers to understand the essential values of the NL – TN canal, not only historically as strategic, economic, ecological and urban landscape morphology but also in the contemporary times as a potential economic and ecological landscape for urban development. Moreover, the research contributed to identifying the existing river landscape elements and values in order to re-integrate them into the practical projects of renovation of the river. Its invisible and potential values were also clearly determined with certain suggestions for some questions related to the urban design and management.

The proposal of a characteristic landscape design structure in preservation - restoration and development of landscape areas for urban design and urban planning of a section of the river with landscape values is composed of three groups of fundamental elements: Culture - Social - Heritage - Historical; Trade - Crafts - Agriculture (local economy); and Environment - Housing - Ecological that shape the identity of the characteristic landscapes structure. These urban landscape elements should be considered appropriately for the future phases of the restoration of the NL – TN canal and that might be referenced for other river landscape design and planning.

## ACKNOWLEDGMENT

This work belongs to the project in 2020 funded by Ho Chi Minh City University of Technology and Education, Vietnam.

## REFERENCES

- [1] Jean, B. (1993). La société au miroir du fleuve. Actes de Colloque International, Lyon, pp. 13-16.
- [2] Jacqueline, B-G. (1995). DOI : 10.4000/geocarrefour.8001; pp. 274. Source: <http://journals.openedition.org/geocarrefour/8001>;
- [3] Duc Minh Quan THAI NGUYEN, The establishment and development of Saigon port urban (XVII- XIX century) (Sự hình thành và phát triển của đô thị cảng Sài Gòn (TK XVI-XIX)), Research topics; pp. 20, 21, 31.  
<https://sites.google.com/site/quankhoasu/su-hinh-thanh-va-phat-trien-cua-dho-thi-cang-sai-gon-tk-xvii-xix?tmpl=%2Fsystem%2Fapp%2Ftemplates%2Fprint%2F&showPrintDialog=1;9/16/2019>
- [4] Huu Thang TRAN, Ba Cuong NGUYEN, Vài nét về kênh Nhiêu Lộc – Thị Nghè xưa và nay, About Nhiêu Loc - Thị Nghè canal in past and present.  
<http://baotang.hcmussh.edu.vn/Resources/Docs/SubDomain/baotang/K%C3%AAnh%20Nh%C3%AAu%20L%E1%BB%99c%20-%20Th%E1%BB%8B%20Ngh%C3%A8.pdf>; p.1, 20/10/2019
- [5, 6, 7] Pierre, G. (1936). Les paysans du delta tonkinois. Etude de géographie humaine, Publication de l'Ecole française d'ExtrêmeOrient, Les Editions d'Art et d'Histoire ; pp. 226, 233, 238.
- [8] Xavier, B. Paul, C. (2011). Etudier une commune : Paysages, Territoires, Populations, Sociétés. Armand Colin ; p. 19.
- [9] A. Dauphin MEUNIER. (1965), Le Cambodge de Sihanouk ou de la difficulté d'être neutre, Paris ; pp. 56.
- [10] Citadel of Saigon. Source: [https://en.wikipedia.org/wiki/Citadel\\_of\\_Saigon](https://en.wikipedia.org/wiki/Citadel_of_Saigon)
- [11] Quy Don LE. (1776), "Phủ Biên Tạp Lục", Duy Anh DAO edited. Information Culture Publisher.
- [12] <https://dantri.com.vn/xa-hoi/ao-moi-cho-kenh-nhiu-loc-thi-nghe-1328809785.htm>; 06/02/2012
- [13] MAK, M. and T. Ng. (2005).The art and science of Feng-shui - a study on architects' perception. Building and Environment 40; pp. 429.
- [14] (1999).Source: Maps of Trịnh Hoài Đức [1765-1825], Gia Định Citadel "Gia Định Thành thông chí", Traduction VSH., Education Edition; pp. 33.
- [15] <http://gallica.bnf.fr/ark:/12148/btv1b53062212t.r=Saigon.langEN>. Source: Bibliothèque nationale de France, département Cartes et plans, GE C-3950; 18/02/2014
- [16] Kevin. L. (1998). L'image de la cité. Paris, Dunod
- [17] Philippe, P. (1980). Éléments d'analyse urbaine. Original provenant de l'Université du Michigan, Editions des Archives d'Architecture Moderne.
- [18] Alain, R. (2004).Morphologie urbaine: géographie, aménagement et architecture de la ville. Armand Colin, Paris.
- [19] Jean-Luc, A. (2008). Analyse spatiale, cartographie et histoire urbaine. Marseille, Éditions Parenthèses/MMSH, coll. "Parcours méditerranéens"; 240 p.

# An Investigation on Efficiency of Magnetic Assisted Generator

Vu-Lan Nguyen  
Department of Mechatronics  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
lannv@hcmute.edu.vn

Chang-Ren Chen  
Green Energy Technology  
Research Center  
Kun Shan University  
Tainan, Taiwan  
crchen@mail.ksu.edu.tw

Bo-Jun Zheng  
Green Energy Technology  
Research Center  
Kun Shan University  
Tainan, Taiwan  
a0929106739@gmail.com

Huann-Ming Chou  
Green Energy Technology Research Center,  
Kun Shan University  
Tainan, Taiwan  
hmchou@mail.ksu.edu.tw

Kuo-Chen Yang  
Green Energy Technology  
Research Center  
Kun Shan University  
Tainan, Taiwan  
ai030821@gmail.com

**Abstract**— The generator developed in this paper is an axial magnetic field generator type. Variation of magnetic flux on its coils generates forces on the rotor. Therefore, this design is called Magnetic assisted generator (MAG). In this study, eight parameters were considered, including (i) magnet type, (ii) coil-magnet distance, (iii) number of magnets, (iv) copper wire diameter, (v) number of coil loops, (vi) capacitance, (vii) number of capacitors in series, and (viii) bridge rectifier. To conduct experimental test, electronic loads, tachometers and oscilloscopes were used to measure the output voltage, rotary speed and alternating current waveforms of the MAG. A DC motor was mounted to and directly drove the rotor of the MAG at a speed of around 1500 rpm while the stator of the MAG was connected to an electronic load to measure the output voltage waveforms and the generated power. Finally, based on the test results, the optimal parameter values for design parameters were figured out.

**Keywords**—Magnetic boosting generator, axial generator, electrical generator

## I. INTRODUCTION

In addition to raising energy saving awareness through education and development of alternative energy resources, renewable energy, etc. to slow down the fossil fuel depletion, it is necessary to increase energy consumption efficiency of devices or energy generation efficiency of generators in general. Particularly in the field of electricity generation, the improvement for efficiency of electrical generators related directly to the electro-magnetic structures and specification parameters of the generator. The Lenz's law, as demonstrated in Fig.1, states that when a changing magnetic field caused by a moving magnet enters and cuts a closed circuit coil, an electric current is induced in the coil, which in turn generates a magnetic field opposing the initial magnetic field (Fig.1a). In other words, the newly generated magnetic field has the same polarity with the magnet, thus makes a repulsive force on the magnet. When the magnet leaves the coil, the magnetic field of the coil and the magnet have different polarities so they attract each other and cause the magnet to stagnate (Fig.1b). The same explanation could be inferred when the polarity of the magnet is changed (Fig.1c). If the magnet is mounted on a rotor, these above-mentioned phenomena are

the reasons why the rotor could be accelerated or decelerated. Basing on such working principles, various structures of motors have been developed for different applications together with relevant solutions to improve working efficiency have been presented [1-4].

The working principle of a magnetic boosted generator or a magnetic assisted generator (MAG) is different as shown in Fig.2. One MAG consists of three elements, namely non-magnetic materials, clockwise winding copper coils and S-pole magnets whose flux cuts the wires of the coils. The magnet attached on the rotor is set to cut the coil with the S-pole end. When the magnet enters and cuts the coil, the generated magnetic field in the coil has different polarity from that of the magnet.

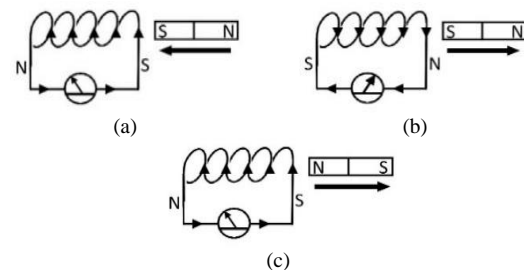


Fig. 1. Diagram of the Lenz's law

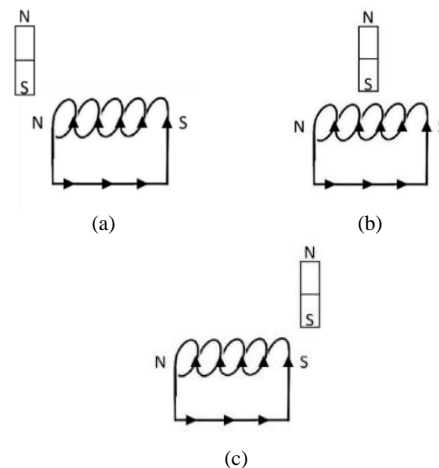


Fig. 2. Diagram of magnetic boosting principle

Therefore, the rotor is sucked in or pushed out by the magnetic field. According to the principle structure of MAG, the magnetic field generated by the wire interacts with the magnet to rotate the rotor. This research investigates the performance of a newly developed MAG which has been awarded an invention patent in Taiwan since 2019 [6]. In addition, the measured parameters were treated by Taguchi's method [7] to figure out their optimal value set.

## II. EQUIPMENT STRUCTURE AND EXPERIMENT PROCESS

The Magnetic assisted generator comprises a magnetic base, a magnetic booster module, and a power generating module (Fig.3). The magnetic base comprises a disc body and a plurality of magnets. The disc body can be driven by power to rotate towards a predetermined direction. The magnetic boosting module, which together with the power generating module is arranged around the outer edge of the disc body and connected in series, comprises a third magnetic pole and a fourth magnetic pole and forms a first magnetic boosting line and a second assisted magnetic line that are kept with an interval, wherein the third magnetic pole is adjacent to the disc body, and the fourth magnetic pole is far away from the disc body. The power generating module comprises a fifth magnetic pole and a sixth magnetic pole and forms a power generating magnetic line, wherein the fifth magnetic pole and the sixth magnetic pole of the power generating module are kept at the same distance to the disc body.

Since the Magnetic assisted generator in this study produces alternating current, a bridge rectifier is used to convert the AC current to DC one. To further stabilize the output voltage, a voltage regulator IC is required. Also, a set of capacitors is connected in parallel with an AC/DC electronic load to filter the output voltage of the generator. The output on the AC/DC electronic load is observed by an oscilloscope so that the capacitance effect and current waveform can be judged. A motor is used to drive the generator of the magnetic booster generator at a fixed speed. After the no-load experiment test, the best parameters of the generator are selected, and the optimization test is carried out again with both no-load and load tests. In this study, eight parameters were considered, including (i) magnet type, (ii) coil-magnet distance, (iii) number of magnets, (iv) copper wire diameter, (v) number of coil loops, (vi) capacitance, (vii) number of capacitors in series, and (viii) bridge rectifier.

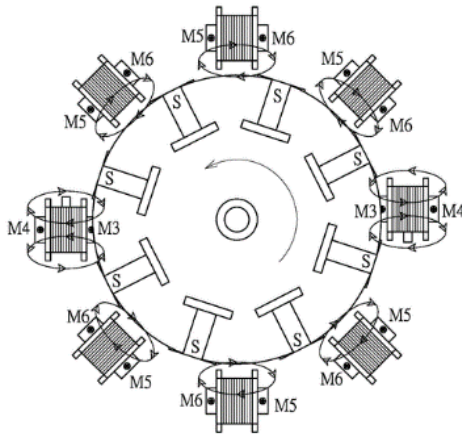


Fig. 3. Principle diagram of the magnetic boosting generator

## III. RESULTS AND ANALYSES

In this research, experiments were conducted in order to find out the power generation performance of the new MAG in relations to variation of 8 different parameters either separately or simultaneously. The selected variation range of each parameter is shown in Table 1. After conducting Taguchi's method, necessary sets of parameters were selected to do experiments. Table 2 shows test results for 18 sets of experimental parameters. Each experiment turn was done with some related parameter(s) changed while all other parameters were set and kept unchanged as shown in Table 2. Table 3 shows measured values of MAG output voltage at different experimental conditions which reveal influences of the parameters on the output power.

TABLE I. EXPERIMENT PARAMETER SETUP RANGE

Item	Selected type/value
Magnet type (A)	N35 or N48 (neodymium iron boron magnets)
Magnet and coil distance (B)	5 mm – 20 mm
Number of magnetic boosters (C)	2 – 8
Copper wire diameter (D)	0.3 mm – 0.5 mm
Number of coil loops (E)	700 - 1000
Filtering capacitor(s) (F)	470 $\mu$ F – 1000 $\mu$ F (200V)
Number of capacitors in series (G)	2 - 6
Bridge rectifier (H)	0.5A – 3A (400V)

TABLE II. EXPERIMENT CONDITION SETUPS

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
	1	2	3	4	5	6	7	8
L1	N35	5mm	2	0.3	700	200V - 470 $\mu$ F	2	0.5A - 400V
L2	N35	5mm	6	0.4	900	200V - 680 $\mu$ F	4	2A - 600V
L3	N35	5mm	8	0.5	1000	200V - 1000 $\mu$ F	6	3A - 1000V
L4	N35	10mm	2	0.3	900	200V - 680 $\mu$ F	6	3A - 1000V
L5	N35	10mm	6	0.4	1000	200V - 1000 $\mu$ F	2	0.5A - 400V
L6	N35	10mm	8	0.5	700	200V - 470 $\mu$ F	4	2A - 600V
L7	N35	20mm	2	0.4	700	200V - 1000 $\mu$ F	4	3A - 1000V
L8	N35	20mm	6	0.5	900	200V - 470 $\mu$ F	6	0.5A - 400V
L9	N35	20mm	8	0.3	1000	200V - 680 $\mu$ F	2	2A - 600V
L10	N48	5mm	2	0.5	1000	200V - 680 $\mu$ F	4	0.5A - 400V
L11	N48	5mm	6	0.3	700	200V - 1000 $\mu$ F	6	2A - 600V
L12	N48	5mm	8	0.4	900	200V - 470 $\mu$ F	2	3A - 1000V
L13	N48	10mm	2	0.4	1000	200V - 470 $\mu$ F	6	2A - 600V
L14	N48	10mm	6	0.5	700	200V - 68 $\mu$ F	2	3A - 1000V
L15	N48	10mm	8	0.3	900	200V - 1000 $\mu$ F	4	0.5A - 400V
L16	N48	20mm	2	0.5	900	200V - 1000 $\mu$ F	2	2A - 600V
L17	N48	20mm	6	0.3	1000	200V - 470 $\mu$ F	4	3A - 1000V
L18	N48	20mm	8	0.4	700	200V - 680 $\mu$ F	6	0.5A - 400V



TABLE III. EXPERIMENT PARAMETER SETUP RANGE

	Measured output voltage (V) at 07 test turns						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
L1	35.06	35.21	35.05	35.19	35.22	35.24	35.13
L2	21.70	21.97	21.79	21.97	21.79	21.91	21.97
L3	36.63	36.46	36.23	36.50	36.65	36.27	36.46
L4	20.64	20.25	20.67	20.29	20.58	20.18	20.61
L5	21.95	22.24	22.14	22.20	22.32	22.30	22.47
L6	19.30	19.37	19.33	19.35	19.37	19.34	19.30
L7	9.22	9.23	9.24	9.22	9.22	9.21	9.21
L8	11.54	11.47	11.52	11.49	11.46	11.47	11.50
L9	4.59	4.56	4.57	4.56	4.55	4.57	4.56
L10	55.95	55.57	56.90	56.10	55.31	55.93	55.61
L11	21.39	21.46	21.42	21.42	21.43	21.42	21.41
L12	35.01	34.90	34.90	34.79	34.89	34.86	34.88
L13	47.61	48.01	47.97	47.96	48.50	47.99	48.19
L14	14.64	14.84	15.02	15.26	15.15	15.29	15.28
L15	20.50	20.48	20.50	20.59	20.65	20.70	20.66
L16	14.09	14.42	14.45	14.37	14.56	14.65	14.72
L17	9.26	9.27	9.25	9.27	9.21	9.20	9.20
L18	5.44	5.16	5.15	5.14	5.14	5.15	5.17

According to the collected data in Table 3, it could be explained, for instant, that measured output voltage in unload-mode of L1 condition was recorded between 35.05 (V) and 35.24 (V) when the generator speed was maintained around 1500 RPM. Experiment system was set as shown in Fig.4 and the measured values were shown on display screen as in Fig.5. Figures 6 to 9 are examples for the measured waveforms of output voltages at conditions L9, L10, L16 and L18, respectively.

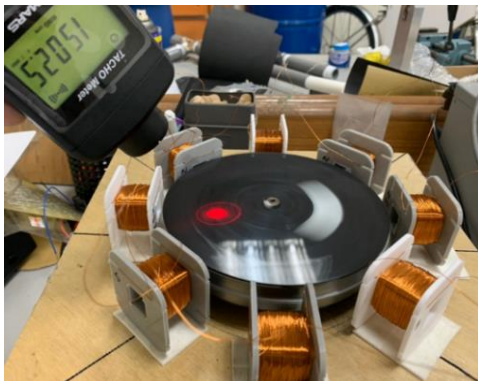


Fig. 4. Experiment setup for the newly designed MAG

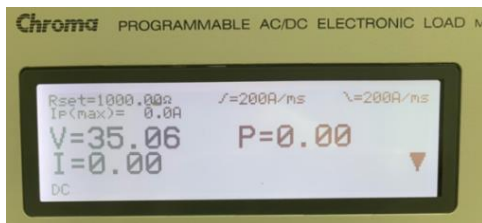


Fig. 5. Measured output values

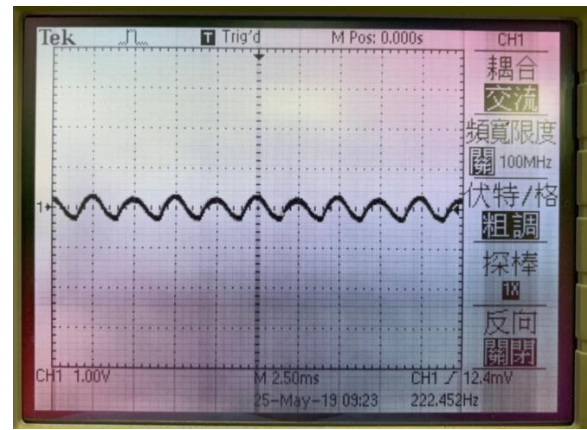


Fig. 6. Measured output voltage waveform on oscilloscope for L9

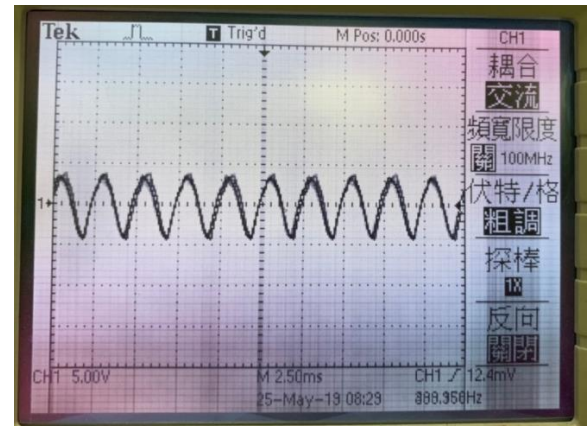


Fig. 7. Measured output voltage waveform on oscilloscope for L10

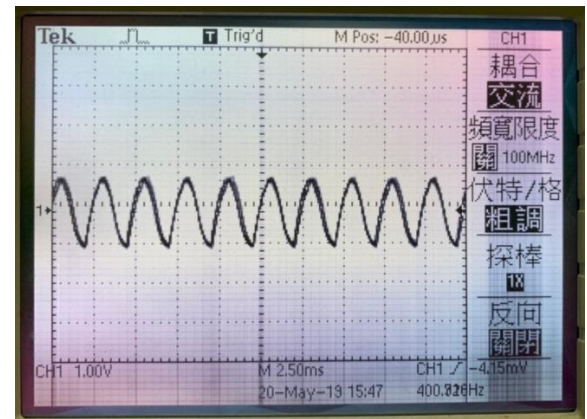


Fig. 8. Measured output voltage waveform on oscilloscope for L16

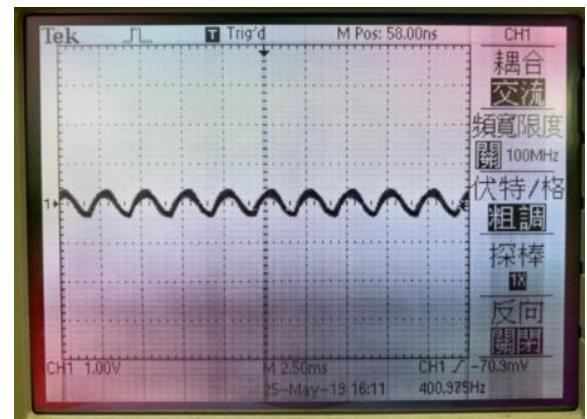


Fig. 9. Measured output voltage waveform on oscilloscope for L18



It could be seen that in some cases, the output voltage magnitude in the 18 parameter conditions varied from around average values of 4.57V (for L9) to 55.95V (for L10). This significant difference was caused mainly by the combined influences of magnet type, magnet-coil distance, number of coils and wire diameter. The individual effect of each parameter could be inferred from Table 2 and Table 3. When magnet type was changed from N35 to the stronger one N48, it was clear that the stronger magnetic power led to higher output result. Another method to increase magnetic interaction flux in the MAG model is to reduce the distance between magnets and coils. It is interesting to find out that when magnet-coil distance reduced from 10mm to 5mm, there would be obviously an increase in output voltage magnitude since the magnetic flux was stronger. Yet, there was a reduction of output voltage magnitude from 21.87V (L2) to 22.23V (L5). This happened because there was also a change of number of coil wire loops from 900 (L2) to 1000 (L5) which caused the larger reduction of induced power. Thus the increase of magnetic flux thank to the distance attenuation in these cases is not enough to compensate the reduction due to the decrease in wire loops. On the other hand, variation of the number of boosting coils made similar effects. As long as the necessary capacitance attached and the power specifications of bridge rectifier used were large enough, the output voltage wave was fully formed and filtered without saturation. After conducting the Taguchi's model analysis, the following parameter set (Table 4) has been achieved as the optimal one for the designed MAG sample.

TABLE IV. OPTIMAL EXPERIMENT PARAMETER SET

Item	Selected type/value
Magnet type (A)	N48 (neodymium iron boron magnets)
Magnet and coil distance (B)	5 mm
Number of magnetic boosters (C)	2
Copper wire diameter (D)	0.5 mm
Number of coil loops (E)	1000
Filtering capacitor(s) (F)	680 $\mu$ F - 200V
Number of capacitors in series (G)	4
Bridge rectifier (H)	0.5A - 400V

#### IV. CONCLUSIONS

With the goal of maximizing voltage and stabilizing output waveforms, the authors propose the following modification for six parameters. First, as the magnet is one of the control factors, between neodymium iron boron magnets N35 and N48, N48 magnet should be used since it has stronger magnetic flux than N35, thus, leads to stronger induced current in the coils. Secondly, as the rotor magnet and the coil are placed perpendicular to each other, it is obvious that the closer the magnet is to the wire, the stronger the magnetic field

and the ability to induce current are. For this reason, it is predicted that if the MAG prototype is manufactured more precisely with coil – magnet gap smaller than 5mm, the achieved output power will be more. Thirdly, increasing magnetic boosters leads to lower output voltages. Although a part of the power generation capacity is sacrificed, power generation effectiveness increases in magnetic assisted generators compared to other traditional generators. Forthly, copper wire diameter of 0.5 mm gives better output power than 0.4 mm and 0.3 mm diameter copper wires because the internal resistance value of the bigger wire is lower. Also, it has been found in the experiment that when more coils or more coil loops are applied, more power is produced. However, it is important to notice that the increase of coil size or number, the mechanical structure design of the generator must be adjusted in a suitable way. Accordingly, there exists a maximum size of the coil size and number in MAG design. Fifthly, it is well known that for capacitors which are connected in series, the equivalent total capacitance value is reversely proportional to the number of the used capacitors. And if a large electric current flows through a small capacitance value, the waveform is unstable and overlaps. Meanwhile, if a small electric current flows through a large capacitance value, the current is filtered by the capacitor and presents a DC waveform. Therefore, a relevant capacitance level must be calculated for the generator in specific circumstances. Finally, when using a bridge rectifier to convert AC to DC, although the waveform after rectification appears as a DC wave, a voltage regulator IC should still necessarily be added to stabilize the output electricity.

#### ACKNOWLEDGMENT

This paper is a part of project No. 109-N-270-EDU-T-048 of Higher Education Deep Cultivation Project of Technical Colleges – Part 2 funded by Taiwan Ministry of Education and Ministry of Science and Technology. The authors would like to deeply thank the Ministries, Green Energy Technology Research Center of Kun Shan University, Taiwan and Ho Chi Minh City University of Technology and Education, Vietnam.

#### REFERENCES

- [1] A. E. Aliasand and Dr. F.T. Josh, "Selection of Motor for an Electric Vehicle: A Review", *Materials Today: Proceedings*, vol. 24, pp 1804–1815, 2020.
- [2] T. A.C. Maiaa, O. A. Fariac, J. E.M. Barros, M. P. Porto, B. J. C. Filho, "Test and simulation of an electric generator driven by a micro-turbine" *Elec. Pow. Sys. Research*, vol. 147, pp. 224-232, June 2017.
- [3] L. I. Farfan-Cabrera, "Tribology of electric vehicles: A review of critical components, current stateand future improvement trends", *Tribology International*, vol. 138, pp.473-486, 2019.
- [4] H. Y. Liu, Y. R. Peng, " Infinite power source generator with multiple magnetic energies", Patent Certificate No.: M548390, website: [https://twpat3.tipo.gov.tw/tipowoc/tipowekm?!!FR\\_M548390](https://twpat3.tipo.gov.tw/tipowoc/tipowekm?!!FR_M548390), valid from 26/09/2016 till 25/09/2026,
- [5] C. R. Chen, Y. Z. Cai, C. C. Lin, J. C. Wang, W. C. Chen, " Magnetic-assisted power generator", Patent Certificate No.: M564862, website: [https://twpat3.tipo.gov.tw/tipowoc/tipowekm?!!FR\\_M564862](https://twpat3.tipo.gov.tw/tipowoc/tipowekm?!!FR_M564862), valid from 22/03/2018 till 21/03/2028.
- [6] C. R. Chen, Y. Z. Cai, C. C. Lin, J. C. Wang, W. C. Chen, Patent Certificate No.: I652883, website: [https://twpat3.tipo.gov.tw/tipowoc/tipowekm?!!FR\\_I652883](https://twpat3.tipo.gov.tw/tipowoc/tipowekm?!!FR_I652883), valid from 22/03/2018 till 21/03/2038.

# Microstructure and Hardness of Borided Layer on SKD61 Steel

Nga Thi-Hong Pham

Faculty of Mechanical Engineering  
HCMC University of Technology and  
Education (HCMUTE)  
Thu Duc-71307, HCM City, Vietnam  
hongnga@hcmute.edu.vn

Long Nhut-Phi Nguyen

Faculty of Mechanical Engineering  
HCMC University of Technology and  
Education (HCMUTE)  
Thu Duc-71307, HCM City, Vietnam  
longnnp@hcmute.edu.vn

The-San Tran

Faculty of Mechanical Engineering  
HCMC University of Technology and  
Education (HCMUTE)  
Thu Duc-71307, HCM City, Vietnam  
santtt@hcmute.edu.vn

**Abstract—** In this study, mechanical properties of boride formed on AISI H13 steel substrate were investigated. The steel has high chromium content and have a widespread use in the engineering application. Boriding treatment was carried out in a mixture of 30% SiC + 70% Borax at 900-920°C for 6 hours. The properties of the boride layer are also achieved by micro hardness methods. The results show that the microstructure after boriding, microstructure has two outer layers are FeB, the middle layer is Fe<sub>2</sub>B. After boriding, the hardness is much higher than the original state (52 HRC), is twice that of the SKD61 steel's hardness. Thus, the study has proposed a suitable heat treatment process to increase the mechanical properties for SKD61 steel to meet the working requirements.

**Keywords—**Boriding, SKD61 steel, Microstructure, Hardness, Boriding mixture

## I. INTRODUCTION

Hot work tool steels has high chromium content and is commonly used engineering material, in which SKD61 steel is the tool materials that are used almost exclusively on extrusion dies as well as for tools for hot pressing of copper alloys and steel forging. It characterized by high strength and ductility, good tempering resistance, moderate cost and also well suited and established for surface treatments [1,2].

One of the surface treatments is boriding which well developed and widely used in industry to produce extremely hard and wear resistant surface layer on metallic substrates. Diffusion boriding is a thermo-chemical treatment that permits boride layers of good performance properties to be produced on steels. The borided steels exhibit high hardness, high wear resistance and improved corrosion resistance [3,4]. Boron atoms can diffuse into ferrous alloys due to their relatively small size and very mobile nature. Boriding can be applied to a wide range of steel alloys including carbon steel, low alloy-steel, tool steel, and stainless steel. In particular, boriding can increase the corrosion resistance of low alloy steel to sulfuric, phosphoric, and hydrochloric acids etc.

Research of George K. Kariofillis [5] showed that H13 hot-worked steel after boriding has better corrosion resistance in H<sub>2</sub>SO<sub>4</sub> and H<sub>3</sub>PO<sub>4</sub> than substrate steel. Ali Günen [6] studied AISI H13 steel that was borided with B<sub>4</sub>C and NaBF<sub>4</sub> powder mixture at 800, 900 and 1000°C for 2, 4 and 6 hours, the results showed that AISI H13 steel after boriding, corrosion resistance was 2-4 times higher than original AISI H13 steel. This shows the potential to use AISI H13 steel for marine applications as an alternative to more expensive martensitic and duplex stainless steels.

## II. MATERIALS AND METHODS

SKD61 steel in the supply state of sheets has dimensions of 20x50x25 mm (Fig. 1).

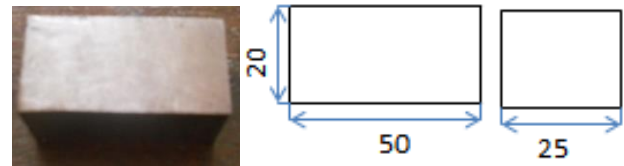


Fig. 1. Test sample

Laboratory materials use borax powder, SiC powder with ratio of 30% SiC + 70% Borax (Fig. 2).



a) Borax powder

b) SiC powder

Fig. 2. Laboratory materials

Laboratory equipments include resistance furnace, ram furnace, graphite container and coal furnace (Fig. 3). After fabrication of the samples, the samples were observed microstructure on the OM and measured hardness on Rockwell machines (Fig. 4).



a) Resistance furnace



b) Tempering furnace



c) Graphite container



d) Coal furnace

Fig. 3. Laboratory equipments



Fig. 4. Equipments for observing the microstructure and hardness measurement

### Experimental procedures

**Annealing:** Heat the oven to 870°C to distribute temperature uniformly inside the furnace, arrange the sample in the container, fill the charcoal to cover up the sample, close and seal the container with clay, place the sample container in the furnace when the furnace reaches the temperature 870°C and keep the heat for 1.5 hours, lower the furnace temperature to 720°C and keep the heat for 2 hours, turn off the furnace and wait until the furnace reaches about 300-400°C, take the container out to cool in air (Fig. 5).

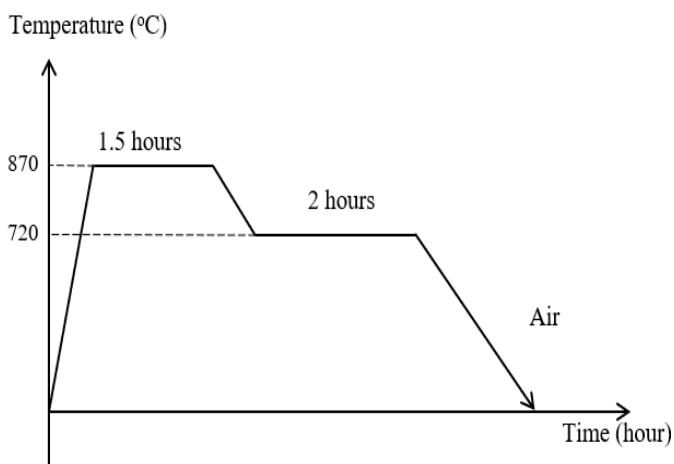


Fig. 5. Steel annealing scheme



a) Container



b) Clay lid sealed the container



c) SKD61 steel samples after annealing

Fig. 6. Boring treatment prepare

**Boring treatment:** Mix 30% SiC + 70% Borax in the container, turn on the furnace to reach about 500°C, put the container (SiC + Borax) into the furnace, after the mixture flows out, put the sample in the container and turn on the furnace to 900-920°C, keep the heat of the container for 6 hours, turn off the furnace and take out the container to cool down (Fig. 6 and Fig. 7).

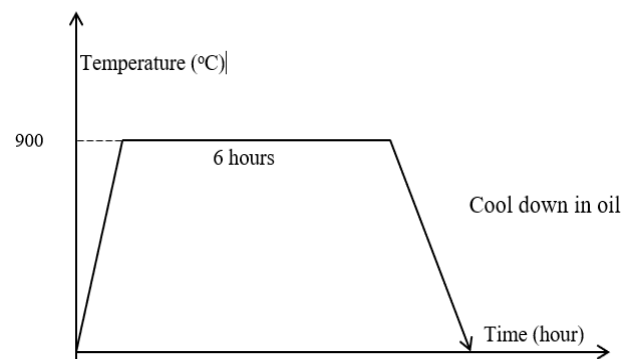


Fig. 7. Boring treatment scheme

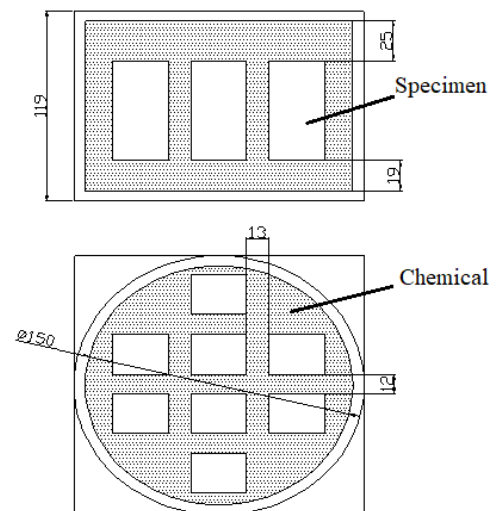


Fig. 8. Placement of SKD61 steel samples into the container



**Hardening:** Arrange the sample in the container, turn on the furnace to reach 1050°C, place the sample container in the furnace, heat the sample container for 25 minutes, turn off the furnace and take out the sample container and cool down in oil.

**Tempering:** Arrange the sample in the container, turn on the furnace to reach 650°C, place the sample container in the furnace, heat the sample container for 2 hours, turn off the oven, take out the sample container to cool down.

### III. RESULTS AND DISCUSSION

#### A. Microstructure

##### The microstructure of SKD61 steel

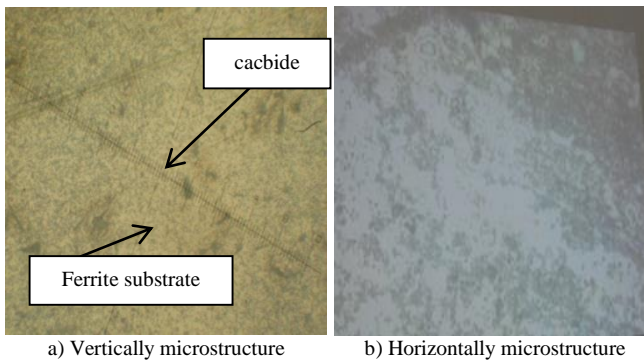


Fig. 9. The original state of the microstructure of SKD61 steel (400X magnification)

Fig. 9 presents the microstructure of SKD61 steel. Based on Fig. 9 it can be seen that the grains in vertically microstructure are uniformly distributed and smaller than that in horizontally microstructure. The microstructure has the Ferrite substrate structure and spherical carbide distributed in it.

##### Microstructure of SKD61 sample after annealing

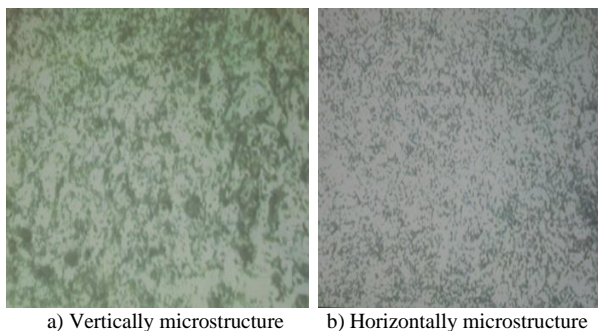


Fig. 10. Microstructure of SKD61 steel after annealing (400X magnification)

Fig. 10 illustrates the microstructure of SKD61 steel after annealing. After annealing, the grain size is more uniform and more evenly distributed. Based on the Fe-C state diagram, because SKD61 steel is the Hypo-eutectoid steel, when heated at temperatures higher than 727°C, austenitic and ferrite microstructure occur, and if the steel heating process continues, there will be the dissolution of ferrite into austenite, when the temperature reaches to 911°C, the process ends and all types of steel will have the same microstructure is fine-grained austenite. When heated

to a temperature of 930-950°C, austenite grains grow very quickly and can produce large grains that exceed the nature of large grains.

##### Microstructure of SKD61 sample after boriding

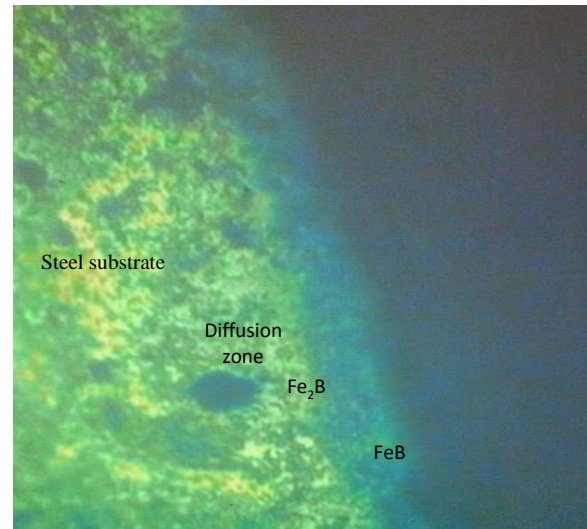


Fig. 11. Microstructure of SKP61 steel after boriding treatment (400X magnification)

Fig. 10 shows the microstructure of SKP61 steel after boriding. From the picture, the boride layer has excellent bonding with the steel substrate, with a uniform thickness of the absorbed layer. After boriding, the microstructure has two outer layers of FeB, the middle layer is Fe<sub>2</sub>B and the inner layer is the substrate. Research by I.Uslu and George K [4,5] showed that the boride layer can form monophase Fe<sub>2</sub>B or double-phase FeB + Fe<sub>2</sub>B on steel substrate surface. In boriding conditions for a long time, it often forms a double-phase. In the case of forming a double-phase, the top layer of the layer is the FeB phase, followed by the Fe<sub>2</sub>B phase, then to the area in contact with the steel base.

Liquid boriding layer is quickly formed but the layer is uneven. For liquid boriding treatment, the process of controlling the boride atomic on the boriding sample surface is pretty difficult because during the process of boriding, the concentration of boride in the molten salt bath will decrease. Therefore, it is necessary to determine the amount of additional salt after each absorbent or to change the temperature and boriding time to have the unchanged boriding microstructure.

##### Microstructure of SKD61 sample after hardening and tempering

Fig. 12 presents the microstructure after hardening and tempering. It can be seen that the grain is smaller, finer, and more uniformly distributed than the original grain. When temper steel reaches to 400°C, the ferrite and cementite appear, during this period, cementite begins to form grain and ferrite changes to multi-edge grain form ferrite. As the temperature continues to rise steadily, the phenomenon of agglomeration and spheroidizing of cementite molecules will gradually grow into a spherical shape. When it reaches the temperature of 500-650°C, the mixture of ferrite and

cementite in grain form is formed or xocbite microstructure is formed. At that time, the formed grains are small smooth and evenly, have good synthesis properties.

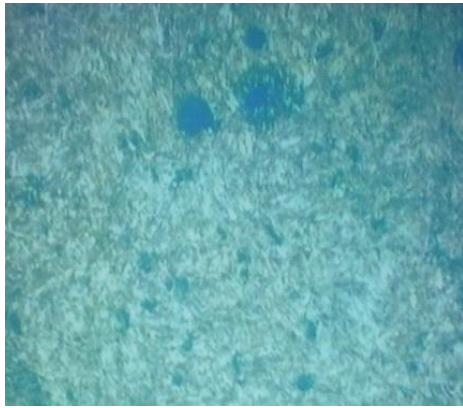


Fig. 12. Microstructure after hardening and tempering (400X magnification)

### B. Hardness of SKD61 steel

TABLE I. RESULT OF HARDNESS TESTING OF SKD61 STEEL (HRC)

Times	Original	After annealing	After boriding	After hardening and tempering
1	24.5	17.5	29	51
2	26	14	28	53,5
3	24.5	16.5	28.5	52
4	24.5	16	27,5	54
5	26	16	27	49
<b>Average (HRC)</b>	<b>25</b>	<b>16</b>	<b>28</b>	<b>52</b>

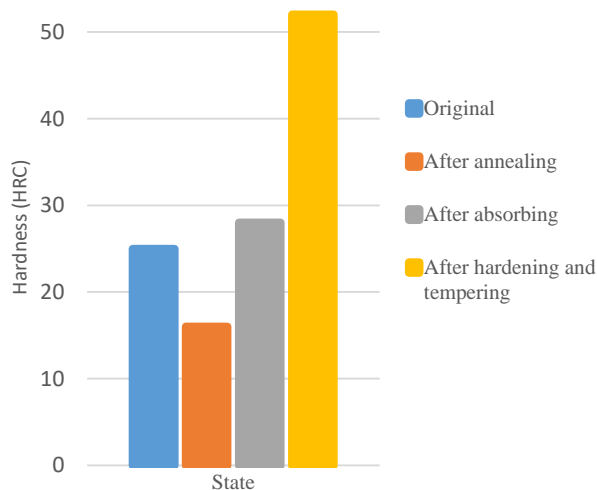


Fig. 13. The hardness of SKD61 steel after annealing, after boriding and after hardening and tempering

Table I and Fig. 13 present the hardness of SKD61 steel after annealing, after boriding and after hardening and tempering. After annealing the hardness decreases rapidly by about 9 HRC compared to the SKD61 steel. After boriding, hardness is 28 HRC, higher than the SKD61 steel. Because after boriding, the boride elements diffuse into the austenitic substrate to form a hard boride layer on the steel surface. Hardness after hardening and tempering has increased significantly by about 52 HRC, which is 2 times

higher than the SKD61 steel. After boriding, the result showed that the mechanical properties are much higher than the SKD61 steel. Boriding allows subsequent heat-treatment processes without reducing the boriding layer properties.

### CONCLUSIONS

- After boriding treatment, microstructure has two outer layers are FeB, the middle layer is Fe<sub>2</sub>B.
- Hardness after boriding is higher than the hardness of the SKD61 steel; after hardening and tempering hardness is 2 times higher than the hardness of SKD61 steel.

For SKD61 steel, it has found an appropriate heat treatment process to increase the mechanical properties of the dies to meet the working requirements: annealing steel at 870°C for 1.5 hours and 720°C for 2 hours, cooling in the air, then boriding at 900°C and keep the heat for 6 hours, hardening at 1050°C for 25 minutes, tempering at 650°C for 2 hours.

### ACKNOWLEDGMENT

We acknowledge HCMC University of Technology and Education, and Material Testing Laboratory (HCMUTE). With their enthusiastic support, we had access to the laboratory and equipment to conduct this research.

### REFERENCES

- [1] G. Kartal, S. Timur, O.L. Eryilmaz, A. Erdemir Influence of process duration on structure and chemistry of borided low carbon steel, vol. 205, 2010.
- [2] Campos-Silva, M. Ortiz-Domínguez, O. Bravo-Bárceñas, et al. Formation and kinetics of FeB/Fe<sub>2</sub>B layers and diffusion zone at the surface of AISI 316 borided steels, vol. 205, 2010.
- [3] Sukru Taktak. Some mechanical properties of borided AISI H13 and 304 steels. Materials & Design, vol. 28(6), 2007
- [4] Uslu, H. Comert, M. Ipek, et al. A comparison of borides formed on AISI 1040 and AISI P20 steels. Materials & Design, vol. 28(6), pp. 1819-1826, 2007.
- [5] George K. Kariofillis, Grigoris E. Kiourtsidis, Dimitrios N. Tsipas. Corrosion behavior of borided AISI H13 hot work steel. Surface and Coatings Technology, vol. 201(1-2), pp. 19-24, 2006.
- [6] Günen Ali, Karahan İsmail Hakkı, Karakaş Mustafa Serdar, et al. Properties and Corrosion Resistance of AISI H13 Hot-Work Tool Steel with Borided B<sub>4</sub>C Powders, Metals and Materials International 2019 / 8, DOI: 10.1007/s12540-019-00421-0
- [7] K. Genel. Boriding kinetics of H13 steel. Vacuum, vol. 80(5), pp. 451-457, 2006.
- [8] Krelling A. P., Milan J. C. G., da Costa C. E. Tribological behaviour of borided H13 steel with different boriding agents. Surface Engineering, vol. 31(8), 2015.
- [9] Wu, Xiao Chun, Peng, Wen Yi, Min, Yongan. Study on Thermal Fatigue Behavior of Boride Layer of H13 Steel. Materials Science Forum, vol. 475-479, pp. 249-252, 2005.
- [10] Boonplook, Yossapong, Juijerm, Patiphan. Prediction of Boride Thickness on Tool Steels AISI D2 and AISI H13 Using Boriding Kinetics. Advanced Materials Research, vol. 931-932, pp. 296-300, 2014.
- [11] López-Perrusquia N., Campos-Silva I., Martínez-Trinidad J., et al. Evaluation of Brittle Layers Obtained by Boriding on AISI H13 Steels. Advanced Materials Research, vol. 65, pp. 47-52, 2009.



# An Experimental Study on The Performance of An Air Conditioning System using CO<sub>2</sub> Refrigerant with The Actual Power Input of 440W

Thanhtrung Dang  
Department of Thermal Engineering  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
trungdang@hcmute.edu.vn

Tronghieu Nguyen  
Faculty of Mechanical Engineering  
HCMC University of Technology and Education  
Ho Chi Minh City, Vietnam  
hieunt@hcmute.edu.vn

**Abstract**—This study presented an experiment on a transcritical CO<sub>2</sub> air conditioning system using a CO<sub>2</sub> compressor with the actual power input of 440 W. Relationships such as evaporation pressure, actual power input, air temperature difference, inlet temperature of expansion, cold room temperature, heat transfer rate, COP were mentioned. At the gas cooler pressure of 78 bar and the evaporation pressure of 50 bar, the COP is 7.9 (with the isentropic compression) for the CO<sub>2</sub> side. For the airside, the COP increases from 2.75 to 4.07 when the evaporator pressure increases from 46 to 50 bar. So, at the evaporation pressure of 50 bar and the cooler pressure of 78 bar, the system should be operated. The COP for the airside is lower than that obtained from the COP for the refrigerant side. In addition, the subcooling method in this study will make the air conditioning system to operate at a lower gas cooler pressure, which led to a more secure and durable system as well as reducing the compression energy

**Keywords**—heat transfer, power input, air conditioning, CO<sub>2</sub> refrigerant, subcooling, COP

## I. INTRODUCTION

In the context of energy-saving and environmental protection, air conditioning systems using CO<sub>2</sub> refrigerants and compact heat exchangers are interesting topics for scientists. Regarding air conditioners using CO<sub>2</sub>, the supercritical and transcritical cycles were studied in [1 - 4]. However, Heo et al. [1] investigated the supercritical Brayton cycle using an isothermal compressor; they did not mention the expansion valve refrigeration cycle as well as the COP (coefficient of performance). In [2], up to 1 MPa high-pressure reduction was achieved with about a maximum of 5% and an average 1% COP losses. Yu et al. [3] evaluated an automobile air conditioning system using a CO<sub>2</sub>-propane mixture as a refrigerant. Under the same compressor speed, the system COP reaches the highest at 60% of CO<sub>2</sub> mass fraction. However, the above studies were tested at high cooler pressure and the main thermodynamic points did not indicate clearly. The two-stage or cascade cycle using CO<sub>2</sub> was presented in [5, 6]. The two-stage compression and the intercooling process are significant in increasing COP for these systems. Comparing a cascade cycle and a single-stage cycle with mechanical subcooling, the best performing levels would be the CO<sub>2</sub> refrigeration cycle with mechanical subcooling. Modeling and numerical simulation of compressors for supercritical carbon dioxide were done in [7 - 9]. Modeling results were presented for the regenerator, the piston rod

diameter, the size of the adiabatic dead volumes, and the working fluid leaks. Numerical CFD simulations were performed to optimize the 3D design of the impellers and of the stators. However, the disagreement in [9] between the numerical simulation and the experimental data might be due to the modeling limitations and the attribution of the subcritical region observed in the contour plot of the static pressure. Investigations on CO<sub>2</sub> expansion in ejectors were done in [10 - 14]. However, they did not research for expansion valves and the main thermodynamic points did not also indicate clearly. Investigations on microchannel heat exchangers and the CO<sub>2</sub> air conditioning cycle were done by Dang et al. [15 - 17]. However, the results in [15, 16] only dealt with the single-phase water; they did not mention CO<sub>2</sub> as the working fluid. The subcooling process of a transcritical CO<sub>2</sub> air conditioning cycle working with a microchannel evaporator was indicated. However, the experimental system in [17] is completely different in this study. The cold room of the system in [17] is very small; with its volume is about 2.3 m<sup>3</sup>. In addition, the results in [17] only mentioned for the refrigerant side, not for the airside.

From the literature reviews above, the studies did not deal with the main thermodynamic points of CO<sub>2</sub> air conditioning systems in more detail and the COP values were not high enough. In addition, the results did not mention the airside in the CO<sub>2</sub> refrigeration cycle. So, it is important to experiment with a CO<sub>2</sub> air conditioning system to determine the physical parameters for the airside. In the following sections, an experimental study of a CO<sub>2</sub> air conditioning system will be done with transcritical mode. The CO<sub>2</sub> compressor with the power input of 440W will be installed in the system.

## II. METHODOLOGY

### A. Mathematical equations

The governing equations were used to analyze the thermodynamic parameters of the CO<sub>2</sub> air conditioning system for the refrigerant side:

The heat transfer rate of gas cooler was calculated as:

$$q_{2-3} = h_2 - h_3 \quad (1)$$

The theory power input was determined using:

$$w_{1-2} = h_2 - h_1 \quad (2)$$

The isenthalpic process was presented by:

$$h_3 = h_4 \quad (3)$$

The heat transfer rate of evaporator was calculated as:

$$q_{4-1} = h_1 - h_4 \quad (4)$$

The COP of the cycle was quantified by:

$$COP_r = \frac{q_{4-1}}{w_{1-2}} \quad (5)$$

For the air side, the equations were used as follow

The heat transfer rate of gas cooler was used as:

$$Q_c = V_c \rho c_p (t_{co} - t_{ci}) \quad (6)$$

The heat transfer rate of evaporator was used as:

$$Q_e = V_e \rho c_p (t_{ei} - t_{eo}) + m_w h_l \quad (7)$$

The COP of the cycle (ignoring the heat transfer rate of subcooling tube) was quantified by:

$$COP_a = \frac{Q_e}{P} \quad (8)$$

The pressure optimization parameter:

$$\alpha = \frac{p_c}{p_e} \quad (9)$$

The pressure optimization parameter:

$$\beta = \frac{p_c}{p_{crit}} \quad (10)$$

(With  $p_{crit} = 73.77$  bar for CO<sub>2</sub>)

where  $m$  is mass flow rate,  $V$  is volumetric flow rate,  $h$  is enthalpy,  $p$  is pressure,  $t$  is temperature (subscripts  $c$  stands for cooler,  $e$  stands for evaporator,  $ci$  stands for cooler inlet,  $co$  stands for cooler outlet,  $ei$  stands for evaporator inlet,  $eo$  stands for evaporator outlet,  $r$  stands for refrigerant,  $a$  stands for air,  $crit$  stands for critical,  $l$  stands for latent heat,  $w$  stands for condensate water, 1-2 stands for compression process, 2-3 stands for cooling process at the cooler, and 4-1 stands for evaporation process),  $\rho$  is density,  $c_p$  is specific heat at constant pressure,  $q$  is heat transfer rate for the refrigerant side,  $Q$  is heat transfer rate for the air side,  $w$  is theory power input, and  $P$  is actual power input.

### B. Experimental setup

The test loop for the CO<sub>2</sub> air conditioning system is indicated in Fig. 1. This system has main components: a CO<sub>2</sub> compressor, a gas cooler, a throttle valve, and a microchannel evaporator. After getting out of the evaporator, the CO<sub>2</sub> refrigerant enters the compressor, and then it is compressed to a higher pressure corresponding higher temperature state. Unlike the study in [17] using the DORIN compressor, this study used the SANDEN CO<sub>2</sub> compressor with the power input of 440 W. The superheated CO<sub>2</sub> is sent to a gas cooler where it is cooled by the atmospheric air. The gas cooler was used is a fin and tube heat exchanger with the capacity of 3200

W. Running out of the gas cooler, the liquid refrigerant is sent to the throttle valve where it becomes wet saturated vapor and its pressure is decreased dramatically. The evaporator has microchannels; the design cooling capacity for this evaporator is 2600 W. The cooler and evaporator were tested with the hydraulic testing method. In this study, the CO<sub>2</sub> air conditioning cycle was done with the transcritical mode.

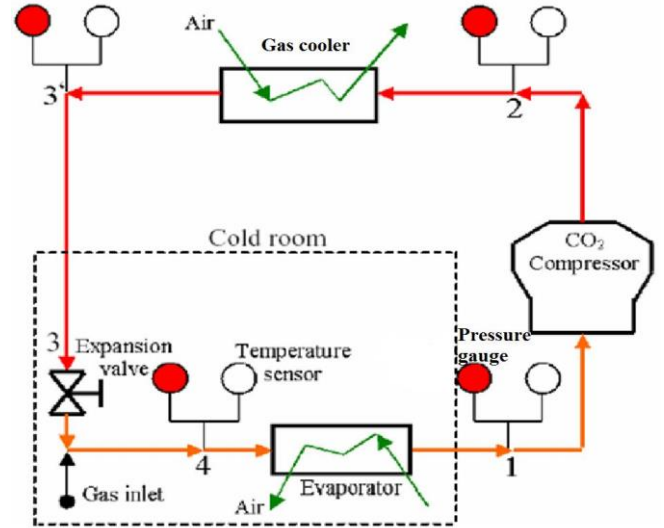


Fig. 1. The test loop for CO<sub>2</sub> air conditioning.

Unlike the experiments in [17] also, a subcooler was not used; instead, this system was installed with a throttle valve in the cold room. This is different from traditional air conditioners that install the throttle valves at the outdoor unit. The refrigerant mixture after the expansion has a lower temperature than the temperature of the cold room and the CO<sub>2</sub> is routed through the microchannels in the evaporator to cool air. The saturated vapor from the evaporator is routed back to the CO<sub>2</sub> compressor to complete a cycle. For the main thermodynamic points of this cycle, pressure gauges, pressure sensors, and temperature sensors were installed to record the parameters. Besides, to increase the data accuracy of the thermodynamic nodes, an infrared thermometer and a thermal camera were used to achieve data. The mass flow rate of condensate water at the evaporator was obtained by the electronic balance.

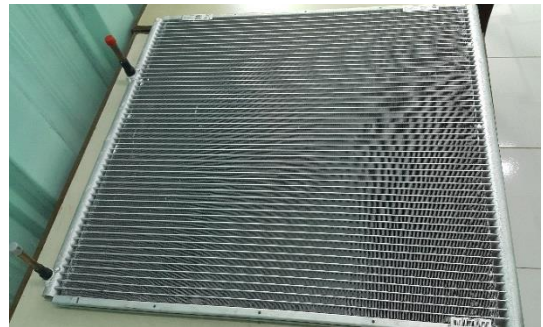


Fig. 2. A photo of the microchannel evaporator.

The gas cooler and evaporator were tested with the hydraulic testing method. A photo of the microchannel evaporator is shown in Fig. 2. The material for this heat exchanger is aluminum, used as a substrate with the thermal conductivity of 237 W/(m°C). The evaporator has six passes with 29 microchannels and each microchannel is rectangular

in shape. The total heat transfer area of this microchannel evaporator is 2.5 m<sup>2</sup>. In this study, the CO<sub>2</sub> air conditioning cycle was done with transcritical mode. The CO<sub>2</sub> air conditioning system will be used to cool a room with the 6.8 m in length, 2.4 m in width, and 3.5 m in height.

TABLE I. ACCURACIES AND RANGES OF TESTING APPARATUSES.

Device	Range	Accuracy
Thermocouple	0 - 100oC	±0.2oC
Thermal camera Fluke	-20 - 250oC	±2%
Pressure gauge	0 - 100 kgf/cm2	±1FS
Pressure sensor	0 - 100 bar	±0.5FS
Digital power meter	0 - 20kW	±0.5%
Anemometer	0 -45 m/s	±3%
Air humidity meter Tenmars TM-181	1 - 99 RH	±2.5% RH
Clamp meter	0 - 200A	±1.5%



Fig. 3. A photo of the experiment system

The types of equipment used for the experiments such as thermocouples, thermostat, Infrared thermometer, thermal camera, pressure gauge, pressure sensor, electronic balance, anemometer, and voltmeter ammeter power meter. A photo of the experimental system is shown in Fig. 3. Accuracies and ranges of testing apparatuses are listed in Table I

### III. RESULT AND DISCUSSION

For experiments carried out in this study, the parameters were fixed as listed in Table II. Under the average ambient temperature of 29.5°C, Table III shows the thermodynamic parameters of the cycle for main points. The subscripts from 1 - 4 in Table III are corresponding with points in the test loop in Fig. 1. It is observed that the temperature at point t3 is equal to the ambient temperature, it is due to the throttle valve was

installed in the air-conditioned room (cold room) and near the evaporator. In addition, the gas cooler in this study was selected with a high heat transfer area, so the outlet CO<sub>2</sub> temperature of cooler is slightly higher than the ambient temperature (31.2°C to compare with 29.5°C). This is a new ideal to increase the COP of the CO<sub>2</sub> air conditioning system. With conventional air conditioners, the throttle valves are installed in the outdoor (consists of compressor and condenser). This conventional system will be lost the amount of energy from the throttle valve to the cold room wall. Besides, the inlet temperature of expansion is still high. So expansion efficiency for these systems is not high, especially for CO<sub>2</sub> refrigerant.

TABLE II. THE FIXED PARAMETER

Parameters	Values	Unit
Volumetric flow rate of air for the evaporator	0.2	m3/s
Volumetric flow rate of air for the gas cooler	0.4	m3/s
Ambient temperature	29.5	°C
Voltage	220	V

TABLE III. THE EXPERIMENTAL POINTS OF THE CYCLE

P (W)	U (V)	p1 (bar)	t1 (oC)	p2 (bar)	t2 (oC)
420	220	49	18.5	79	57
p3' (bar)	t3' (oC)	p3 (bar)	t3 (oC)	p4 (bar)	t4 (oC)
78	31.2	78	29.6	50	14.4

(where P is actual power input, U is Voltage, p is pressure, and t is temperature)

The data in Table 3 were drawn on the p-h diagram of CO<sub>2</sub> using the EES software as shown in Fig. 4 (EES is an acronym for Engineering Equation Solver). From Fig. 4, at the gas cooler pressure of 78 bar and the evaporator pressure of 50 bar, the power input is 18.9 kJ/kg (with the isentropic compression), the cooling capacity is 149.8 kJ/kg, resulting to the COP is 7.9. These results were determined for the CO<sub>2</sub> side. They will be used to compare with the parameters from the airside in the next section

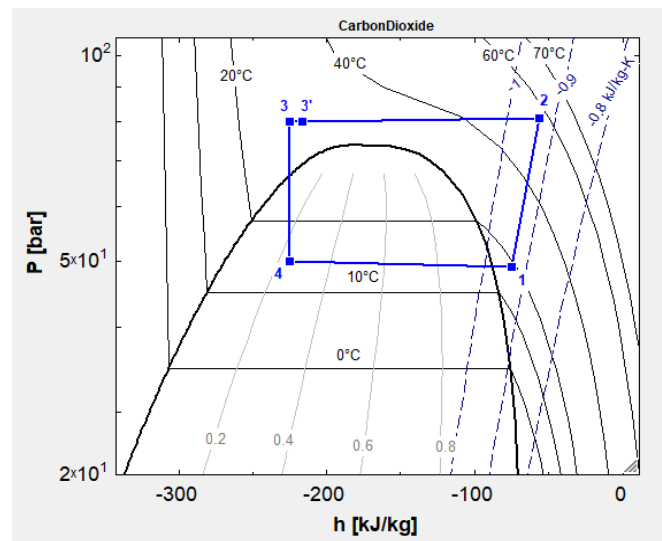


Fig. 4. The thermodynamic points of the cycle on p-h diagram.

A relationship between the gas cooler pressure, the factor ( $\alpha$ ), and the inlet temperature of expansion are shown in Fig. 5. The factor ( $\alpha$ ) is the ratio between the gas cooler pressure and the evaporation pressure. In this study, when the gas cooler pressure increases from 75 bar to 81 bar, the factor ( $\alpha$ ) and the inlet temperature of expansion increase. The inlet temperature of expansion is equal to the ambient temperature because the throttle valve has been placed in the cold room after the refrigerant leaves the gas cooler, it continues to release heat to the cold room outside and the cold room inside. It is noted that the dimensions of the cold room were mentioned in the experimental setup section of this study. With the released heat from the gas cooler to the cold room outside, this is a benefit over the method of placing a throttle at the outside (at the outdoor unit). With the release heat from the cold room wall to the throttle valve, it has reduced the refrigerant temperature to approximately the room temperature to be cooled (less than the ambient temperature). As the cold room air temperature dropped, it caused the throttle inlet temperature to drop further. In many experiments in this study from 2015 to the present, the results showed that no more the subcooling heat (the subcooling tube section from the cold room wall to the throttle valve) affects the cooling load in the cold room.

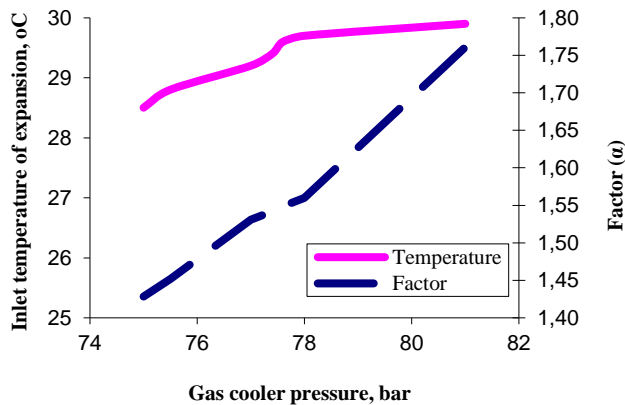


Fig. 5. Gas cooler pressure vs. temperature and factor ( $\alpha$ )

At the balance state, the system has a pressure value of 58 bar. From the experimental data, when the evaporator pressure increases from 46 to 52 bar, the actual power input obtained from the power meter decrease from 440 to 390 W (corresponding with the current decreases from 2.2 to 2.0 A). It is noted that the actual power input was obtained from the voltmeter ammeter power meter. During the experimental period, the average condensate water at the evaporator is 0.5 kg/h, calculating the average latent heat of 340 W. With the calculation based on the equation (7), the heat transfer rate of the evaporator for airside increases from 1.21 to 1.71 kW when the evaporator pressure increases from 46 to 50 bar; however, the heat transfer rate decreases from 1.71 to 1.47 kW when the evaporator pressure increases from 50 to 52 bar, as shown in Fig. 6. The change of heat transfer rate may be also depended on the inlet temperature of expansion and the air temperature difference (between the inlet air and the outlet air) of the evaporator. For the airside of the evaporator, when the air temperature difference increases from 3.5 to 5.8°C, the heat transfer rate increases from 1.14 kW to 1.71 kW, as shown in Fig. 7

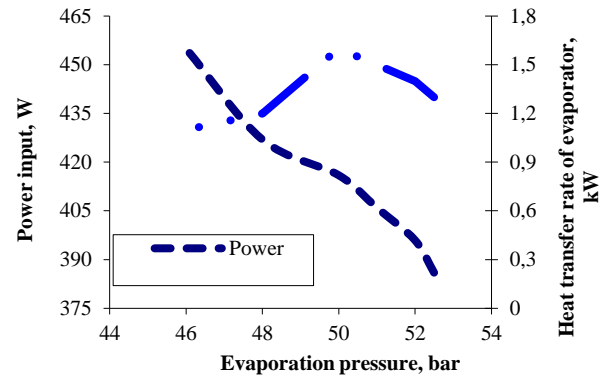


Fig. 6. Evaporation pressure vs. power input and heat transfer rate

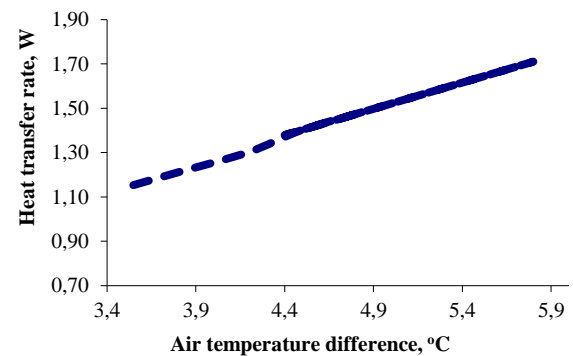


Fig. 7. Heat transfer rate and air temperature difference of the evaporator .

A relationship between the outlet air temperature of the evaporator and the inlet temperature of expansion is shown in Fig. 8. It is observed that the inlet temperature of expansion increases from 28.3°C to 29.7°C, the outlet air temperature of the evaporator also increases from 19.4°C to 21°C. It is leading that the cold room temperature increases from 25.2°C to 26.2°C, as indicated in Fig. 9. The results also affected the heat transfer rate of the evaporator in this cycle.

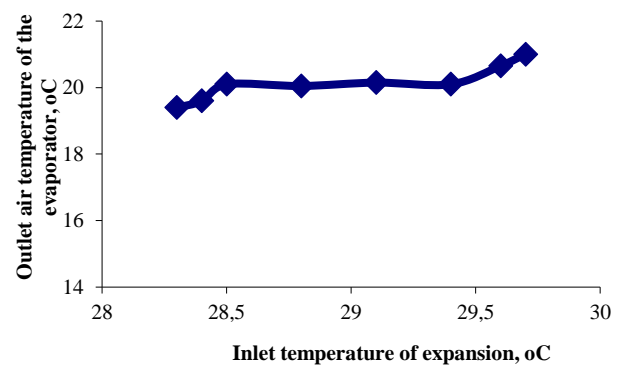


Fig. 8. Outlet air temperature vs. inlet temperature of expansion.



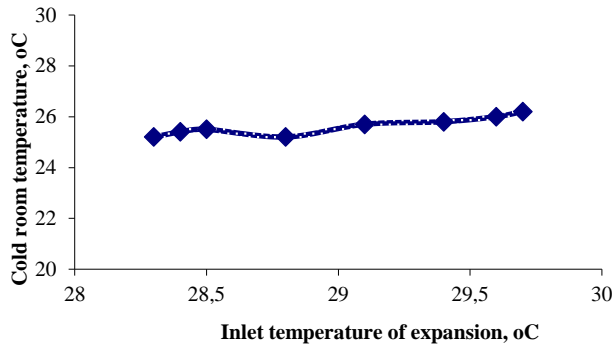


Fig. 9. Cold room temperature vs. inlet temperature of expansion.

From the power meter and heat transfer rate of the evaporator for the airside, the COP for airside is calculated and shown in Fig. 10 as changing the evaporator pressure. When the evaporation pressure increases from 46 to 50 bar, the COP for airside increases from 2.75 to 4.07; however, the evaporation pressure increases from 50 to 52 bar, the COP for airside decreases from 4.07 to 3.67.

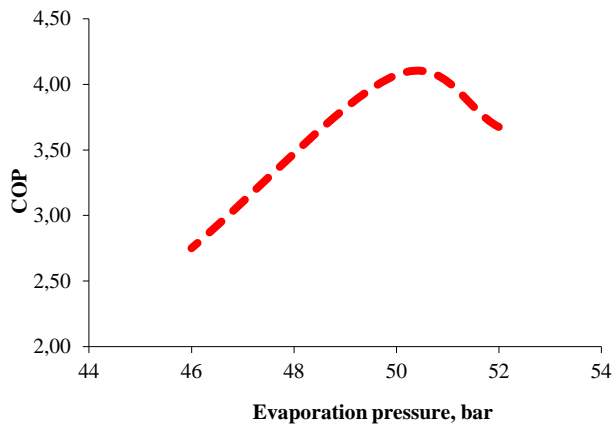


Fig. 10. COP for air side and evaporation pressure

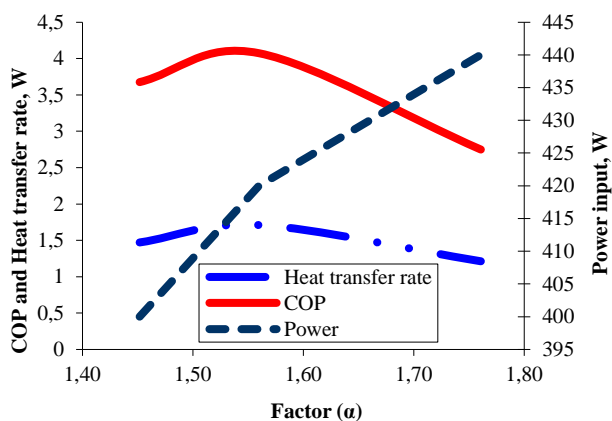


Fig. 11. The factor ( $\alpha$ ) vs. the heat transfer rate of evaporator, the power input, and the COP

A relationship between the factor ( $\alpha$ ), the heat transfer rate of the evaporator, the actual power input, and the COP is shown in Fig. 11. The experimental results show that the increase of the factor ( $\alpha$ ) caused the actual power input

increases. When the factor ( $\alpha$ ) increases from 1.45 to 1.56, the COP for airside increases from 3.67 to 4.07; however, the factor ( $\alpha$ ) increases from 1.56 to 1.76, the COP for airside decreases from 4.07 to 2.75. The COP reaches its maximum value at the factor ( $\alpha$ ) of 1.56, corresponding with the evaporation pressure of 50 bar and the gas cooler pressure of 78 bar. Similar results were found for the factor ( $\beta$ ), as shown in Fig. 12. Here, the factor ( $\beta$ ) is the ratio between the gas cooler pressure and the critical pressure of  $\text{CO}_2$ . In our opinion, the factor ( $\alpha$ ) shows the thermodynamic characteristics of the refrigeration system more clearly than the factor ( $\beta$ ) because the factor ( $\alpha$ ) represents both the gas cooler pressure and the evaporation pressure. The results of the relationship between the COP and evaporation temperature, evaporation pressure, the factor ( $\beta$ ) in this study are similar to those in [18 – 20]. However, the present study has a higher COP and a lower factor ( $\beta$ ) because the results in this study were obtained at higher evaporation pressure and lower gas cooler pressure than those studied in [18 – 20]

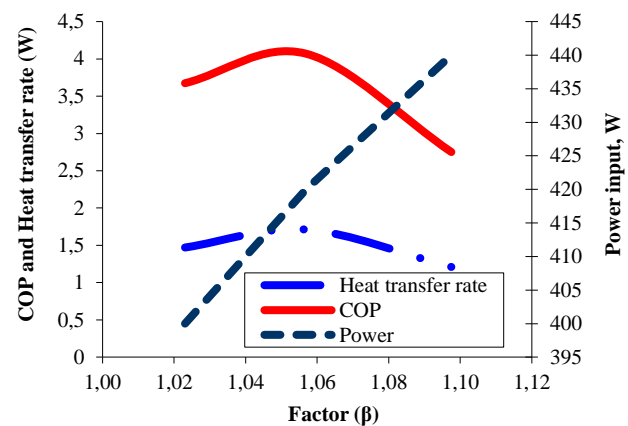


Fig. 12. The factor ( $\beta$ ) vs. the heat transfer rate of evaporator, the power input, and the COP.

From figures 10 - 12, the system should be operated at the evaporator pressure of 50 bar (corresponding to the cooler pressure of 78 bar) to achieve the maximum COP. Comparing with the COP for the refrigerant side, the COP for the airside is lower than that obtained from the COP for the refrigerant side. This deviation is due to the compression efficiency, the heat transfer efficiency of the evaporator, and the experimental errors. From the results in this study as well as the characteristics of  $\text{CO}_2$  refrigerant, the  $\text{CO}_2$  air conditioning cycle is necessary to have a subcooling system. For the air conditioning systems using the Hydrofluorocarbon (HFC) or Hydrochlorofluorocarbon (HCFC) refrigerant, the installation of the throttle valve at the outdoor unit or the indoor unit is not as significant as the  $\text{CO}_2$  air conditioning system. The subcooling method when placing the throttle valve at the indoor unit will make the system simpler than the subcooling methods using the auxiliary cooling system as in [21, 22]. The subcooling method in this study makes the air conditioning system to operate at a lower gas cooler pressure, which led to a more secure and durable system as well as reducing the compressor power.

It is noted that these results have not been published from the literature reviews. The results have made important contributions to the research on  $\text{CO}_2$  air conditioning. It is



more practical in the context of energy-saving and environmental protection.

#### IV. CONCLUSION

An experimental investigation on a CO<sub>2</sub> air conditioning system with the power input of 440 W was done with transcritical mode. Using EES software, the experimental points were plotted on the p-h diagram to indicate the thermodynamic points as well as to calculate the heat transfer rate for the refrigerant side. Based on the experimental results, the conclusions can be summarized as follows:

- At the gas cooler pressure of 78 bar and the evaporation pressure of 50 bar, the power input is 18.9 kJ/kg, the cooling capacity is 149.8 kJ/kg (with the isentropic compression), resulting to the COP is 7.9 for the CO<sub>2</sub> side.
- When the evaporation pressure increases from 46 to 50 bar, the actual power input obtained from the power meter decreases from 440 to 420 W. However, the heat transfer rate of the evaporator for airside increases from 1.21 to 1.71 kW.
- When the air temperature difference increases from 3.5 °C to 5.8°C, the heat transfer rate for the airside of the evaporator increases from 1.14 kW to 1.71 kW.
- When the inlet temperature of expansion increases from 28.3°C to 29.7°C, the outlet air temperature also increases from 19.4°C to 21°C, and the cold room temperature increases from 25.2°C to 26.2°C.
- The COP for airside increases from 2.75 to 4.07 when the evaporation pressure increases from 46 to 50 bar. In this study, the system should be operated at the evaporation pressure of 50 bar and the cooler pressure of 78 bar.
- Comparing with the COP for the refrigerant side, the COP for the airside is lower than that obtained from the COP for the refrigerant side.
- In addition, the subcooling method when placing the throttle valve at the indoor unit will make the air conditioning system to operate at a lower gas cooler pressure, which led to a more secure and durable system as well as reducing the compression energy. The results are important contributions to the research on CO<sub>2</sub> air conditioning cycle.

#### ACKNOWLEDGMENT

The supports of this work by the projects T2021 sponsored by the specific research fields of Ho Chi Minh City University of Technology and Education are deeply appreciated.

#### REFERENCES

- [1] J. Heo, M. Kim, S. Baik, S. Bae, and J. Lee. 2017. Thermodynamic study of supercritical CO<sub>2</sub> Brayton cycle using an isothermal compressor. *Applied Energy* 206, 1118–1130
- [2] L. Shao, Z. Zhang, and C. Zhang. 2018. Constrained optimal high pressure equation of CO<sub>2</sub> transcritical cycle. *Applied Thermal Engineering* 128, 173–178
- [3] B. Yu, D. Wang, C. Liu, F. Jiang, J. Shi, and J. Chen. 2018. Performance improvements evaluation of an automobile air conditioning system using CO<sub>2</sub>-propane mixture as a refrigerant. *International Journal of Refrigeration* 88, 172–181
- [4] Y. Ma, Z. He, X. Peng, and Z. Xing. 2012. Experimental investigation of the discharge valve dynamics in a reciprocating compressor for transcritical CO<sub>2</sub> refrigeration cycle. *Applied Thermal Engineering* 32, 13–21
- [5] I. M. G. Almeida and C. R. F. Barbosa. 2011. Performance analysis of two-stage transcritical refrigeration cycle operating with R744. *Proceedings of COBEM*, pp. 1–10
- [6] L. Nebot-Andrés, ID R. Llopis, D. Sánchez, J. Catalán-Gil and R. Cabello. 2017. CO<sub>2</sub> with mechanical subcooling vs. CO<sub>2</sub> cascade cycles for medium temperature commercial refrigeration applications thermodynamic analysis. *Applied sciences* 7, 1–22
- [7] N. Holaind , G. Bianchi , M. Miol , S. Saravi , S. Tassou, A. Leroux , and H. Jouhara. 2017. Design of radial turbo machinery for supercritical CO<sub>2</sub> systems using theoretical and numerical CFD methodologies. *Energy Procedia* 123, 313–320
- [8] R. Ibsaine , J. Joffroy, and P. Stouffs. 2016. Modelling of a new thermal compressor for supercritical CO<sub>2</sub> heat pump. *Energy* 117, 530–539
- [9] S. Kim, J. Lee, Y. Ahn, J. Lee, Y. Addad, and B. Koc. 2014. CFD investigation of a centrifugal compressor derived from pump technology for supercritical carbon dioxide as a working fluid. *The Journal of Supercritical Fluids* 86, 160–171
- [10] Y. Huai, X. Guo, and Y. Shi. 2017. Experimental study on performance of double-throttling device transcritical CO<sub>2</sub> ejector refrigeration system. *Energy Procedia* 105, 5106 – 5113
- [11] M. Palacz, J. Smolka, W. Kus, A. Fic, Z. Bulinski, A. Nowak, K. Banasiak, and A. Hafner. 2016. CFD-based shape optimisation of a CO<sub>2</sub> two-phase ejector mixing section. *Applied Thermal Engineering* 95, 62–69
- [12] M. Palacz, M. Haida, J. Smolka, A. Nowak, K. Banasiak, and A. Hafner. 2017. HEM and HRM accuracy comparison for the simulation of CO<sub>2</sub> expansion in two-phase ejectors for supermarket refrigeration systems. *Applied Thermal Engineering* 115, 160–169
- [13] M. Palacz, J. Smolka, A. Nowak, K. Banasiak, and A. Hafner. 2017. Shape optimization of a two-phase ejector for CO<sub>2</sub> refrigeration systems. *International journal of refrigeration* 74, 210–221
- [14] Y. Zhu, C. Li, F. Zhang, and P. Jiang. 2017. Comprehensive experimental study on a transcritical CO<sub>2</sub> ejector-expansion refrigeration system. *Energy Conversion and Management* 151, 98–106
- [15] T. Dang and T. Teng. 2011. The effects of configurations on the performance of microchannel counter-flow heat exchangers—An experimental study. *Applied Thermal Engineering* 31, 3946–3955
- [16] T. Dang and T. Teng. 2011. Comparison on the heat transfer and pressure drop of the microchannel and minichannel heat exchangers. *Heat and Mass Transfer* 47, 1311–1322
- [17] T. Dang, H. Vo, H. Le, and H. Nguyen. 2017. An experimental study on subcooling process of a transcritical CO<sub>2</sub> air conditioning cycle working with microchannel evaporator. *Journal of Thermal Engineering* 3, 1505–1514
- [18] X. Liu, R. Fu, Z. Wang, L. Lin, Z. Sun, and X. Li. 2019. Thermodynamic analysis of transcritical CO<sub>2</sub> refrigeration cycle integrated with thermoelectric subcooler and ejector. *Energy Conversion and Management* 188, 354–365
- [19] E. Bellos and C. Tzivanidis. 2019. Enhancing the performance of a CO<sub>2</sub> refrigeration system with the use of an absorption chiller. *International Journal of Refrigeration* 108, 37–52
- [20] E. Bellos and C. Tzivanidis. 2019. A comparative study of CO<sub>2</sub> refrigeration systems. *Energy Conversion and Management: X* 1, 100002
- [21] R. Llopis, L. Nebot-Andrés, S´anchez D., Catal´an-Gil J., Cabello R.. 2018. Subcooling methods for CO<sub>2</sub> refrigeration cycles - A Review. *International Journal of Refrigeration*
- [22] B. Yu, J. Yang, D. Wang, J. Shi, and J. Chen. 2019. An updated review of recent advances on modified technologies in transcritical CO<sub>2</sub> refrigeration cycle. *Energy*.

# Redundant Relay Protection Devices for Power Systems Reliability Model

Andrey Trofimov<sup>1</sup>

<sup>1</sup>*Novosibirsk State Technical University*  
Novosibirsk, Russia  
a.trofimov@corp.nstu.ru

Alexandra Khalyasmaa<sup>2,1</sup>

<sup>2</sup>*Ural federal university*  
Ekaterinburg, Russia  
a.i.khalyasmaa@urfu.ru

<sup>1</sup>*Novosibirsk State Technical University*  
Novosibirsk, Russia  
xalyasmaa@corp.nstu.ru

**Abstract**—The study's purpose was to develop a mathematical model for the redundant RP systems functioning under health poll and to study the influence of RP failures on the RP reliability indicators. The study result was to find the dependence on the initial characteristics of such indicators as the unavailability factor, average time between failures in standby mode, the RP failure probability in external short circuit modes and in damage to the securable unit, the RP mean time effect on the reliability and operation indicators. The model of the RP system operation cycle based on semi-Markov processes theory has been developed. The model allows calculating and forecasting operation and reliability indicators, such as RP system's unavailability factor, recovery frequency, and technical position checks frequency. The above model allowed minding three different types of failures, which allows obtaining the frequencies of these failures.

**Keywords**—*relay protection, redundancy, faults, reliability.*

## I. INTRODUCTION

Ensuring structural reliability and efficiency of relay protection (RP) functioning requires the building of such models that are suitable for theoretical and experimental study of their properties. Mathematical modeling is the most prolate and promising method for studying these systems, which allows researches at the design stage, solving analysis and synthesis problems, predicting the RP systems quality and efficiency, justifying their necessary or optimal structure and correctly interpreting statistical data.

Redundancy is one of the main ways to improve the technical systems reliability. At present, in Russia and abroad, relay protection system redundancy is performed "on actuation" (switching on two or more protection sets for securable unit disconnecting independently of each other according to the "or" scheme). In some European countries, redundancy "on non-actuation" is used, the action of two relay protection system sets according to the "and" scheme (tripping when two protection sets are operated simultaneously). At the most critical facilities, it may be necessary to include three identical protection sets with a two-out-of-three tripping, i.e. action to disconnect the securable unit only when two sets are operated simultaneously and a fault alert is generated when only one protection set was operated.

Reliability in accordance with three types of the relay protection system functions is divided into the reliability upon internal faults, the security of non-operation upon external short circuits (SC) and reliability in regular operating

modes of the electric power system (EPS) - standby modes. In accordance with the standard [1], the security and reliability concepts are defined. Failure to perform the required functions by the protection may consist in unintended operation, for example, a false tripping, or in the RP failure [2,3]. In this case, it is convenient to use the resulting reliability concept, which defines the totality of all the given reliability types and is called the unavailability factor [4].

The use of the above-mentioned relay protection system redundancy schemes has a different effect on the reliability and the security. Ensuring the structural reliability and efficiency of the RP system redundancy operation requires the models suitable for theoretical and experimental study of their properties, depending on the redundancy scheme. The mathematical models of relay protection systems are based on the processes during their operation: failures occurrence and elimination, checking equipment, etc.

The study purpose was to develop a mathematical model for the redundant RP systems functioning under health poll, to study the influence of RP failures on the RP reliability indicators. The study result was to find the dependence on the initial characteristics of such indicators as the unavailability factor, average time between failures in standby mode, the RP failure probability in external short circuit modes and in damage to the securable unit, the RP mean time effect on the reliability and operation indicators.

## II. ELECTRIC POWER SYSTEM RELAY PROTECTION OPERATION

The RP operation consists in the interaction of various events stream with the RP system, such as defects in relaying schemes, internal faults and external short circuits, regular checks and recovery in different states. Usually there is an unambiguous connection between the type of defects in the scheme and the possible type of failure in operation. If there is a defect in the scheme, dangerous from the point of view of RP actuation failures, then in case of damage, a RP failure to trip may occur. In non-redundant schemes by non-actuation, the defect dangerous from the point of view of false tripping can immediately lead to RP false tripping [5].

The RP system reliability is significantly influenced by two types of events: components failures and their performance restoration. The RP system operation is due to the interaction, on the one hand, of events stream, and on the other hand, to the internal structure, the redundancy, the RP system state and parameters.

During the RP system operation events stream arises: short circuit inside and outside the protection; RP system repairs; RP system preventive checks; RP system actuation; RP system failures, etc. These events stream can be divided into two groups: regular and random streams. Most of them are random. Random streams consist of events, the occurrence of which is unknown in advance, these events occur at random times. Regular streams are those in which events occur at predetermined, known times. These usually include streams of RP system preventive checks, periodic test checks streams. In this regard, the problem of simultaneous accounting for the renewal stream random and regular components arises.

### III. RELAY PROTECTION SYSTEM MODELING: METHODOLOGY

A correct solution to the described problem is possible within the theory of Markov processes. To analyze the reliability, we will take the power transmission line protection system, which initiates the protected line's tripping if it is damaged. The protection set's failures flows parameters and the recovery rate are known (Table 1).

TABLE I. INITIAL EVENTS RATES

No.	Title	Sign	Value, 1/year
1	False tripping intensity	$\lambda_{f,A} = \lambda_{f,B} = \lambda_f$	0,006
2	Unnecessary operations intensity	$\lambda_{u,A} = \lambda_{u,B} = \lambda_u$	0,104
3	Failure to trip intensity	$\lambda_{ft,A} = \lambda_{ft,B} = \lambda_{ft}$	0,012
4	Periodical check intensity	$\lambda_p$	365-0,08
5	External short circuit intensity	$v_i$	2,7
6	Internal faults intensity	$v_{un}$	2,45

From the RP system operable state, it is possible to enter disabled state with defects that are dangerous from the point of view of failure to trip or unnecessary operations (latent failures).

After a latent failure occurs, the system is disabled, which can be interrupted in two ways.

- The first method - when a line short circuit occurs, the relay fails to trip, or when external short circuits occur, unnecessary operations occur, after which the relay operable state is restored within a few hours.
- The second method - if a line short circuit or an external short circuit does not occur before the periodic check start, then the latent failure is detected during the regular check at the end of the period  $t_p$ .

Also, from the protection set operable state, you can enter a RP system abnormal operation state in standby mode. RP system abnormal operation state in the absence of a short circuit is called "false tripping" (detectable failure), leading to a RP line false tripping with the failure flow parameter. After detecting a failure with the failure detection intensity, the RP system recovery comes up within several hours.

#### A. State and transition graph

Let us consider a RP model variant with "on actuation" redundancy. Table 2 lists the possible states of the relay protection system with "on actuation" redundancy and transitions between them.

Fig. 1 shows a state and transition graph of the relay protection system with "on actuation" redundancy.

On the graph, the RP system operable state is indicated in green, the RP system operable state with the defects that are

dangerous from the point of view of failure to trip or unnecessary operations - in yellow, RP system failures, types "false tripping", unnecessary operations and failure to trip - in red, RP system periodic checks in operable and disabled states - in gray, RP system recovery state- in blue.

TABLE II. RP SYSTEM STATE WITH "ON ACTUATION" REDUNDANCY

1.OS	RP system operable state consisting of two operable protection set A and B
2.UN <sub>f,A</sub>	false tripping of the set A
3.UN <sub>f,B</sub>	false tripping of the set B
4.O <sub>u,A</sub>	RP system operation with the defects that are dangerous from the point of view of unnecessary operations of the set A
5.O <sub>u,B</sub>	same defect of the set B
6.O <sub>ft,A</sub>	RP system operation with the defects that are dangerous from the point of view of failure to trip of the set A
7.O <sub>ft,B</sub>	same defect of the set B
8.UN <sub>u</sub>	unnecessary operations of the set A or B
9.UN <sub>ft</sub>	failure to trip of the set A or B
10.CD	RP system check, in disabled state
11.CO	RP system check, in operable state
12.R	RP system recovery state

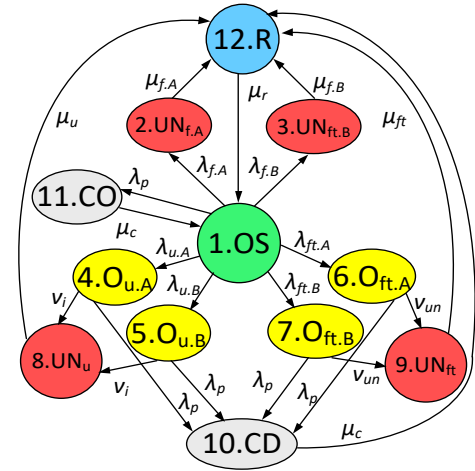


Fig. 1. State and transition graph of the relay protection system with "on actuation" redundancy

In the relay protection system, consisting of two identical protection sets A and B (initiating the protected unit tripping independently of each other according to the "or" scheme), one of the types of failures occurs in random time. Protection sets' A or B false tripping comes out and leads to the RP system failure (transition 1→2, 1→3), which corresponds to states 2, 3 - false tripping. The time the system is in the failure state is several hours, after which the recovery follows the transition to state 12 with the failure detection intensity of set A false tripping -  $\mu_{f,A}$  and set B -  $\mu_{f,B}$  (transition 2→12 и 3→12).

In the defects that are dangerous from the point of view of A or B set unnecessary operations, the relay system changes to disabled state and unlocked state (transitions 1→4 and 1→5). As for the failure - unnecessary operation occurs only when two events are superimposed: a defect in set A or B and an external short circuit with intensity  $v_{int}$  (transitions 4→8 and 5→8). After the RP system failure the recovery follows (transition 8→12).

In the defects that are dangerous from the point of view of the set A or B failures to trip, the relay system changes to the disabled and unlocked state (transitions 1→6 и 1→7). As for the failure to trip it occurs only when two events are superimposed: a defect in set A or B and an internal fault with an intensity  $\lambda_{u,A}$  (transitions 6→9 and 6→9). After a relay protection system failure to trip, the recovery follows (transition 9→12). After the recovery, the relay system returns to an operable state (transition 12→1).

It is accepted that the events leading to the state transition are random and the time before these events is distributed according to the exponential law. Then the transitions between states are characterized by the following constant intensities:  $\lambda_{u,A}$ ,  $\lambda_{u,B}$  – the unnecessary operations intensity of the sets A, B;  $\lambda_{f,A}$ ,  $\lambda_{f,B}$  – the false tripping intensity of the sets A, B;  $\lambda_{ft,A}$ ,  $\lambda_{ft,B}$  – the failures to trip intensity of the sets A, B;  $\lambda_c$  – the check intensity;  $\mu_c$  – the check completion intensity;  $\mu_r$  – the recovery completion intensity (restoration intensity),  $\mu_{f,A}$ ,  $\mu_{f,B}$  – the false tripping detection intensity of the protection sets A, B;  $\mu_u$ ,  $\mu_{ft}$  – the failures detection rates of the type of unnecessary operations and the RP system failure to trip;  $v_i$  – the external short-circuits intensity (outside the protected unit),  $v_{int}$  – the internal faults intensity (at the protected unit).

A specific feature of the system under consideration is the following: the time spent in states 2, 3, 8, 9 is negligible compared to the time spent in states 1, 4, 5, 6, 7. Therefore, the time spent in states 2, 3, 8, 9 can be neglected. Theoretically, this means that the exiting from these states rate is infinitely large, that is,  $\mu_c=\infty$ ,  $\mu_u=\infty$ ,  $\mu_{ft}=\infty$ ,  $\mu_{f,A}=\mu_{f,B}=\infty$ .

The average recovery time is conventionally taken equal to 10 hours. So, in states 4, 5, 6, 7 the RP system is disabled, in these states the system is for a finite time. In other disabled states, as well as in test states 10 and 11, the system is for a negligible time.

#### IV. MATHEMATICAL MODEL

The presented model belongs to the class of analytical models based on the Markov processes theory [6,7,8,9,10]. Since the time spent in states 2, 3, 8, 9, 10, 11 is negligible (theoretically it is equal to zero), then after entering these states there is a “reflection” from these states, which consists in an instant transition to other states.

This determined the mathematical model choice, namely, the transitions between states are described by a semi-Markov process, in which only state changes are recorded. The transitions between states will be described by the probabilities  $p_{ij}$ , which, in contrast to the Markov circuits transition probabilities, will be called the transmission probabilities. The transmission probability  $p_{ij}$  – is the probability that a transition  $i \rightarrow j$  between  $i$  and  $j$  will occur, provided that there is an exit from state  $i$  [8,11]. So, the above prerequisites make it possible to compose an initial model of the redundant RP system functioning based on a semi-Markov process.

The total intensity of exit from states 1, 4, 5, 6, 7:

$$\lambda_{11} = \lambda_p + \lambda_{f,A} + \lambda_{f,B} + \lambda_{u,A} + \lambda_{u,B} + \lambda_{ft,A} + \lambda_{ft,B} = \lambda_p + 2\lambda_f + 2\lambda_u + 2\lambda_{ft};$$

$$\lambda_{44} = \lambda_{55} = \lambda_p + v_i; \quad \lambda_{66} = \lambda_{77} = \lambda_p + v_{int}.$$

Transitions from states 1, 4, 5, 6, 7 to other states are described by the transition probabilities, which are proportional to the corresponding transitions intensities:

$$p_{12} = \lambda_{f,A} / \lambda_{11}; \quad p_{13} = \lambda_{f,B} / \lambda_{11}; \quad p_{14} = \lambda_{u,A} / \lambda_{11}; \quad p_{15} = \lambda_{u,B} / \lambda_{11};$$

$$p_{16} = \lambda_{ft,A} / \lambda_{11}; \quad p_{17} = \lambda_{ft,B} / \lambda_{11}; \quad p_{111} = \lambda_{ft} / \lambda_{11}; \quad p_{48} = v_i / \lambda_{44};$$

$$p_{410} = \lambda_p / \lambda_{44}; \quad p_{58} = v_i / \lambda_{55}; \quad p_{510} = \lambda_p / \lambda_{55}; \quad p_{69} = v_{int} / \lambda_{66};$$

$$p_{610} = \lambda_p / \lambda_{66}; \quad p_{79} = v_{int} / \lambda_{77}; \quad p_{710} = \lambda_p / \lambda_{77}.$$

The transition probability matrix between states has the form:

$$P = \begin{pmatrix} 0 & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} & 0 & 0 & 0 & p_{111} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{48} & 0 & p_{410} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{58} & 0 & p_{510} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{69} & p_{610} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{79} & p_{710} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

It is convenient to represent the system functioning in time in the form of cycles. Let's divide the states set into two subsets according to the phases of operation:

$$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}; \quad V = \{12\}.$$

In the subset  $U$  states, the system is functioning and checked, and in the subset  $V$  state it is recovered. Over time the system is in the subset  $U$  and after being in this subset it changes to the subset  $V$ . Then this  $UV$  cycle is repeated.

Transition probabilities submatrices on subsets  $U$  and  $V$  are:

$$P_{UU} = \begin{pmatrix} 0 & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} & 0 & 0 & 0 & p_{111} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{48} & 0 & p_{410} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{58} & 0 & p_{510} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{69} & p_{610} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_{79} & p_{710} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$P_{VV} = (0).$$

The matrices  $P_{UU}$  and  $P_{VV}$  are used to calculate the matrices of relative frequencies  $N_U$  and  $N_V$  on the subsets  $U$  and  $V$  [8,12]:

$$N_U = \|n_U(i, j)\| = (E - P_{UU})^{-1} \quad (1)$$

$$N_V = \|n_V(i, j)\| = (E - P_{VV})^{-1} \quad (2)$$

where  $E$  – identity matrix of the corresponding order,  $n_U(i, j)$  – average relative frequency of the  $j$  – th state, or mathematical expectation of the entering number into the  $j$  – th state before leaving the subset  $U$  provided that the  $i$  – th state of this subset is initial.

The following meaning is given to the relative frequency:  $n_U(i, j)$  – the mathematical expectation of the entering (intrusions) number in the  $j$  – th state before leaving the subset  $U$ , per one recovery. Since state 1 during the transition  $V \rightarrow U$  is always initial, the relative frequencies of the subset  $U$  states are described only by the first row of the matrix  $N_U$ . The first row of the matrix  $N_U$  has the form:

$$\bar{n}_1 = \|n_U(1, j)\| \quad (3)$$

$$\bar{n}_1 = \frac{1}{\lambda_{11} - \lambda_p} \left( \lambda_{11} \lambda_{1f} \lambda_{1f} \lambda_{1u} \lambda_{1u} \lambda_{1f} \lambda_{1f} \cdot \frac{2 \cdot \lambda_{1u} \cdot \lambda_{1u}}{\lambda_{44}} \frac{2 \cdot \lambda_{1u} \cdot \lambda_{1u}}{\lambda_{66}} \frac{2 \cdot \lambda_{1p} \cdot (\lambda_{1f} \cdot \lambda_{44} + \lambda_{1u} \cdot \lambda_{66})}{\lambda_{44} \cdot \lambda_{66}} \lambda_p \right)$$

Obviously,  $N_V = (1)$ .

To obtain the states and states subsets temporal characteristics, a matrix of mean residence time in the subset  $U$  states is introduced for a single intrusion in the states:

$$\Theta_U = \begin{pmatrix} \frac{1}{\lambda_{11}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_{44}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\lambda_{55}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\lambda_{66}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\lambda_{77}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Then the row of mean residence time in the subsets  $U$  and  $V$  on one cycle has the form:

$$\bar{t}_U = \|t_U(1, j)\| = \bar{n}_1 \cdot \Theta_U \quad (4)$$

$$\bar{t}_U = \frac{1}{\lambda_{11} - \lambda_p} (1 \ 0 \ 0 \ \frac{\lambda_{1u}}{2\lambda_{44}} \ \frac{\lambda_{1u}}{2\lambda_{55}} \ \frac{\lambda_{1f}}{2\lambda_{66}} \ \frac{\lambda_{1f}}{2\lambda_{77}} \ 0 \ 0 \ 0 \ 0) \quad (5)$$

An element  $t_U(1, j)$  – the average time spent in the  $j$  – the state per one recovery. Now it is possible to get a formula for the mean residence time in the subset  $U$ :

$$t_U = \bar{t}_U \cdot \dot{e} = \frac{(v_i + \lambda_p) \cdot (v_{um} + \lambda_p) + 2\lambda_u \cdot (v_{um} + \lambda_p) + 2\lambda_{fi} \cdot (v_i + \lambda_p)}{2 \cdot (v_i + \lambda_p) \cdot (v_{um} + \lambda_p) \cdot (\lambda_u + \lambda_f + \lambda_{fi})} \quad (6)$$

where  $\dot{e}$  – is a column with all elements equal to 1.

The average  $UV$ – cycle time is

$$t_{UV} = t_U + t_V = \frac{(v_i + \lambda_p) \cdot (v_{um} + \lambda_p) + 2\lambda_u \cdot (v_{um} + \lambda_p) + 2\lambda_{fi} \cdot (v_i + \lambda_p)}{2 \cdot (v_i + \lambda_p) \cdot (v_{um} + \lambda_p) \cdot (\lambda_u + \lambda_f + \lambda_{fi})} + t_V$$

where  $t_V$  – is the RP system accepted average recovery time from 2 to 10 hours in accordance with the RP system operational records.

The average cycle frequency, meaning the average number of recoveries per time unit (in this case, one year), is calculated using the formula:

$$\omega_{UV} = 1/t_{UV} \quad (7)$$

The total mean residence time spent in disabled states 4, 5, 6, 7 on one  $UV$ – cycle:

$$t_r = t_U(1, 4) + t_U(1, 5) + t_U(1, 6) + t_U(1, 7) = \frac{\lambda_u(v_{um} + \lambda_p) + \lambda_{fi}(v_i + \lambda_p)}{(v_i + \lambda_p)(v_{um} + \lambda_p)(\lambda_u + \lambda_f + \lambda_{fi})}$$

The unavailability factor calculated in accordance with:

$$K_{un} = \frac{t_r}{t_{UV}} = \frac{t_r / t_U}{1 + t_V / t_U}, \quad (8)$$

where

$$\frac{t_r}{t_U} = \frac{2\lambda_u(v_{um} + \lambda_p) + 2\lambda_{fi}(v_i + \lambda_p)}{(v_i + \lambda_p)(v_{um} + \lambda_p) + 2\lambda_u(v_{um} + \lambda_p) + 2\lambda_{fi}(v_i + \lambda_p)}. \quad (9)$$

The average checks number per cycle is

$$n_n = n_U(1, 10) + n_U(1, 11) \quad (10)$$

The average checks frequency is the average number of checks during the year:

$$\omega_p = n_p \cdot \omega_{UV}. \quad (11)$$

The average time between checks  $t_p$  and accordingly, the checks intensity  $\lambda_p$  are parameters that can be set during operation. Table 4 shows the dependence on these parameters of the average cycle time  $t_{UV}$ , the average cycle frequency  $\omega_{UV}$ , the unavailability factor  $K_{un.f}$ , the checks average number per cycle  $n_p$  and the average checks frequency  $\omega_p$ .

The states and transitions graph of a variant of the RP system with “on non-actuation” redundancy is shown in Fig. 3 with fifteen states, which are given in Table. 3.

TABLE III. RP SYSTEM STATE WITH “ON NON-ACTUATION” REDUNDANCY

1.OS	RP system operable state consisting of two operable protection set A and B
2.O <sub>f.A</sub>	RP system operation in false tripping of the set A
3.O <sub>f.B</sub>	RP system operation in false tripping of the set B
4.O <sub>u.A</sub>	RP system operation with the defects that are dangerous from the point of view of unnecessary operations of the set A
5.O <sub>u.B</sub>	same defect of the set B
6.O <sub>u.AB</sub>	same defect of the set A and B
7.O <sub>fi.A</sub>	RP system operation with the defects that are dangerous from the point of view of failure to trip of the set A
8.O <sub>fi.B</sub>	same defect of the set B
9.O <sub>fi.AB</sub>	same defect of the sets A and B
10.UN <sub>fi</sub>	failure to trip of the sets A and B
11.UN <sub>u</sub>	unnecessary operations of the sets A and B
12.UN <sub>f</sub>	false tripping of the sets A and B
13.CD	RP system check, in disabled state
14.CO	RP system check, in operable state
15.R	RP system recovery state



TABLE IV. RP SYSTEM STATE WITH “ON NON-ACTUATION” REDUNDANCY

$t_p$ (year)	$\lambda_p$ (1/year)	“on actuation” redundancy					“on non-actuation” redundancy				
		$t_{UV}$ (year)	$\omega_{UV}$ (1/year)	$K_{un.f}$	$n_p$	$\omega_p$ (1/year)	$t_{UV}$ (year)	$\omega_{UV}$ (1/year)	$K_{un.f}$	$n_p$	$\omega_p$ (1/year)
1/365	365	4,10	0,2438	0,00063	1496,85	365	4,10	0,2438	0,00067	1504,14	365
1/48	48	4,12	0,2428	0,00456	197,62	48	4,12	0,2427	0,00506	204,32	48
1/12	12	4,16	0,2401	0,01556	49,96	12	4,18	0,2391	0,01989	55,11	12
1/4	4	4,24	0,2357	0,03359	16,96	4	4,34	0,2300	0,05699	20,02	4
1/2	2	4,30	0,2324	0,04729	8,60	2	4,58	0,2179	0,10631	10,49	2
1	1	4,36	0,2294	0,05940	4,36	1	5,04	0,1984	0,18657	5,55	1
2	0,5	4,40	0,2273	0,06813	2,19	0,5	5,85	0,1710	0,29885	3,08	0,5
3	0,333	4,41	0,2264	0,07164	1,47	0,333	6,54	0,1528	0,37379	2,25	0,333
4	0,25	4,42	0,2259	0,07354	1,11	0,25	7,16	0,1396	0,42758	1,83	0,25
5	0,2	4,43	0,2257	0,07472	0,88	0,2	7,71	0,1297	0,46822	1,56	0,2
6	0,167	4,43	0,2255	0,07554	0,74	0,167	8,20	0,1219	0,50013	1,38	0,167
7	0,143	4,44	0,2254	0,07613	0,63	0,143	8,64	0,1156	0,52594	1,24	0,143
8	0,125	4,44	0,2253	0,07658	0,55	0,125	9,05	0,1104	0,54731	1,14	0,125
9	0,111	4,44	0,2252	0,07693	0,49	0,111	9,43	0,1060	0,56535	1,05	0,111
10	0,1	4,44	0,2251	0,07721	0,44	0,1	9,78	0,1023	0,58082	0,98	0,1
11	0,09	4,44	0,2250	0,07745	0,40	0,09	10,10	0,0989	0,59427	0,92	0,09
12	0,08	4,44	0,2249	0,07764	0,37	0,08	10,41	0,0961	0,60609	0,87	0,08

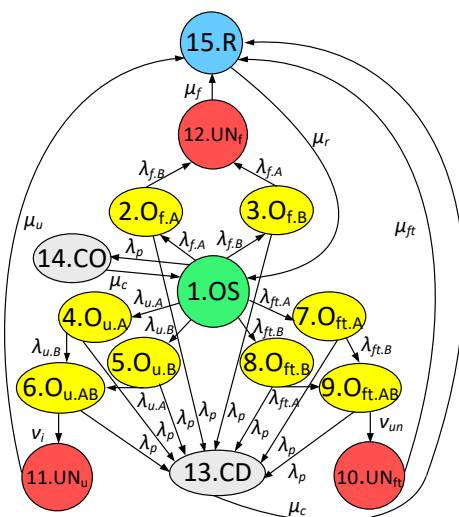


Fig. 2. The states and transitions graph of the RP system with “on non-actuation” redundancy

The total exiting states intensity 1, 2, 3, 4, 5, 6, 7, 8, 9:

$$\begin{aligned} \lambda_{11} &= \lambda_p + \lambda_{f.A} + \lambda_{f.B} + \lambda_{u.A} + \lambda_{u.B} + \lambda_{f_i.A} + \lambda_{f_i.B} = \lambda_p + 2 \cdot \lambda_j + 2 \cdot \lambda_u + 2 \cdot \lambda_{f_i}; \\ \lambda_{22} &= \lambda_p + \lambda_{f.B}; \lambda_{33} = \lambda_p + \lambda_{f.A}; \lambda_{44} = \lambda_p + \lambda_{u.B}; \lambda_{55} = \lambda_p + \lambda_{u.A}; \lambda_{66} \\ &= \lambda_p + \mathbf{v}_j; \lambda_{77} = \lambda_p + \lambda_{f_i.B}; \lambda_{88} = \lambda_p + \lambda_{f_i.A}; \lambda_{99} = \lambda_p + \mathbf{v}_{un}. \end{aligned}$$

Transitions from states 1, 2, 3, 4, 5, 6, 7, 8, 9 to other states are described by the transitions probabilities, which are proportional to the corresponding transitions intensities:

$$\begin{aligned}
P_{12} &= \lambda_{f,A}/\lambda_{11}; P_{13} = \lambda_{f,B}/\lambda_{11}; P_{14} = \lambda_{u,A}/\lambda_{11}; P_{15} = \lambda_{u,B}/\lambda_{11}; \\
P_{17} &= \lambda_{fi,A}/\lambda_{11}; P_{18} = \lambda_{fi,B}/\lambda_{11}; P_{114} = \lambda_{p}/\lambda_{11}; P_{212} = \lambda_{f,B}/\lambda_{22}; \\
P_{213} &= \lambda_{p}/\lambda_{22}; P_{312} = \lambda_{f,A}/\lambda_{33}; P_{313} = \lambda_{p}/\lambda_{33}; P_{46} = \lambda_{u,B}/\lambda_{44}; \\
P_{413} &= \lambda_{p}/\lambda_{44}; P_{56} = \lambda_{u,A}/\lambda_{55}; P_{513} = \lambda_{p}/\lambda_{55}; P_{611} = v/\lambda_{66}; \\
P_{613} &= \lambda_{p}/\lambda_{66}; P_{79} = \lambda_{f,A}/\lambda_{77}; P_{713} = \lambda_{p}/\lambda_{77}; P_{89} = \lambda_{fi,A}/\lambda_{88}; \\
P_{813} &= \lambda_{p}/\lambda_{88}; P_{911} = v/\lambda_{99}; P_{913} = \lambda_{p}/\lambda_{99}.
\end{aligned}$$

The states set based on the RP system lives:

$$U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}; V = \{15\}.$$

Further calculations are performed in accordance with expressions 1-8 for the states and transitions graph for the RP system according to Fig. 3. The results of calculating the operation and RP system redundancy schemes reliability (Fig. 3) indicators are shown in Table 4.

In fig. 4 they build the dependences of the RP system unavailability factor on the average time between checks for RP systems with one protection set in accordance with [12,13], as well as with two protection sets, depending on the redundancy scheme.

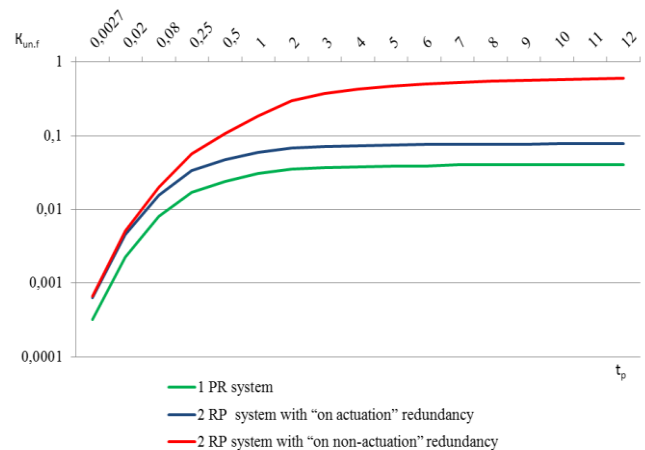


Fig. 3. RP system unavailability factor  $K_{un,f}$  change on the mean time between checks  $t_n$

## V. CONCLUSION

The model of the RP system operation cycle based on semi-Markov processes theory has been developed. The model allows calculating and forecasting operation and reliability indicators, such as RP system's unavailability factor, recovery frequency, and technical position checks frequency. The above model allowed minding three different types of failures, which allows obtaining the frequencies of these failures.

On the given model, it is possible to identify the factors influencing the RP system reliability, and assess the unavailability factor dependence on various influencing factors, as well as assess the alternative redundancy schemes reliability for the RP system.

The model advantage is the possibility of its implementation by computer modeling based on systems such as Mathcad or Matlab.

Using this model, it was found that for small periods between periodic checks up to one month, the unavailability factor does not depend on the RP system redundancy scheme, however, at large values of the checks frequency (from 1 month to 12 years), the difference between the unavailability factor increases by 1.3- 10 times. The average recovery frequency of the RP system with “on actuation” redundancy practically does not depend on the established checks frequency, however, the average checks frequency depends significantly on this factor. The average recovery frequency of the RP system with “on non-actuation” redundancy and the checks frequency average depends on the established checks frequency.

On the given model, the regular checks frequency requirements can be formulated, which provide the required reliability at appropriate operating costs.

#### REFERENCES

- [1] International standard IEC 50 (448) Protection of energy systems. Geneva, 1995.
- [2] Guk Yu.B. Analysis of the reliability of power plants. - L. : Energoatomizdat, 1988. 224 p.
- [3] Shneerson E.M. Digital relay protection. M.: Energoatomizdat, 2007. 549 p.
- [4] GOST R 53480-2009. Reliability in technology. Terms and Definitions.
- [5] Shalin A.I., Reliability and diagnostics of relay protection of power systems. - Novosibirsk: Publishing house of NSTU, 2003. -- 384 p.
- [6] B.P. Zelentsov Analytical modeling of complex probabilistic systems. Modeling of information networks. Proceedings of the Computing Center SB RAS. Series: Computer Science, vol. 1. 1994. P. 144-152
- [7] Kelbert M.Ya., Sukhov Yu.M. Probability and statistics in examples and problems. Vol. II: Markov chains as a starting point for the theory of random processes. M.: MTsNMO 2010.
- [8] Korolyuk V.S. and other Semi-Markov processes and their applications. - Kiev, 1970.
- [9] Yu.K. Belyaev, A. Bogatyrev, V.V. Bolotin and others; Ed. I.A. Ushakova. Reliability of technical systems: Radio and communication, 1985. 608 p.
- [10] Samarskiy A.A. Mathematical modeling. Moscow: Fizmatlit, 2002. . 320 p.
- [11] B.P. Zelentsov Matrix models of system reliability: engineering calculation methods. Novosibirsk: Nauka. 1991. 112 p.
- [12] Zelentsov B.P., Maksimov V.P., Shuvalov V.P. Communication line functioning model under conditions of unreliable control of technical condition. 1991. 112 p.
- [13] Trofimov A.S., Zelentsov B.P. Functioning model of relay protection of power systems. ELECTRIC POWER. Transmission and distribution, 2016, No. 6. P. 110-114.

# Forecasting the Solar Power Plants Generation in a Meteorological Data-Constrained Environment

Andrey Tokarev<sup>1</sup>

<sup>1</sup>Ural Federal University  
Ekaterinburg, Russia  
andreitokarev-1@rambler.ru

Alexandra Khalyasmaa<sup>1,2</sup>

<sup>1</sup>Ural Federal University  
Ekaterinburg, Russia  
a.i.khalyasmaa@urfu.ru

<sup>2</sup>Novosibirsk State Technical University  
Novosibirsk, Russia  
xalyasmaa@corp.nstu.ru

**Abstract**— This paper is devoted to increasing the accuracy of forecasting the solar power plants generation in Russia in a data-constrained environment. Within this study, an analysis of forecasting parameters influence on the solar power plants generation based on the regression analysis methods was made. The study object was a real power system, with the installed capacity of 3915 MW, 330 MW of which is the solar power plants' share. The study also analyzed the change in the average price indice of the balancing market in Russia, depending on solar power plants presence.

**Keywords**— solar power plants, forecasting, generation, regression analysis.

## I. INTRODUCTION

Nowadays, the electric power industry development is directly related to the integration and operation of power generation facilities operating on renewable energy sources (RES). The direct dependence on the RES facilities weather conditions is manifested in the sharply variable nature of the power output. When operating in parallel with the electric power system (power system), such a RES facilities' operation aspect affects the entire power system [1].

First, the influence consists in the complexities of forming the active power reserves [2], the maintenance of which is necessary for reliable and high-quality power supply to consumers. Secondly, the random power output by RES facilities can lead to an excess of the maximum permissible active power flow through a line or group of lines (section) [3], which can cause a static aperiodic stability violation. Therefore, the power systems' system operators, which include RES facilities, should pay special attention to wind power plants (WPP) and solar power plants (SPP) generation forecasting.

The WPP/SPP power generation schedule forecasting can be carried out using various models, based on [4]:

- physical approach [5];
- statistical approach [6];
- machine learning [7];
- combined approach [8].

Each forecasting model has its own accuracy and is selected depending on the required time coverage (forecasting time-frame): short-term, medium-term or long-term.

The basic elements for the correct forecasts for all the above models are the following: the initial data size, their quality, as well as the number of analyzed parameters. But the

RES facilities implementation is often not accompanied by a meteorological stations (meteorological desks) multiplication to form the required amount of data for the analyzed geo-location. This, in turn, can lead to initial data insufficiency and degradation, which characterizes this problem as a problem in data-constrained environment for the correct generation forecast.

## II. ANALYSIS OF THE CURRENT SITUATION AND APPROACHES TO RES GENERATION FORECASTING IN RUSSIA

### A. Approaches to RES generation forecasting

The world experience in managing power systems with a significant share of RES facilities is characterized by the same-type measures to ensure the safe and reliable power systems operation. These measures are investment in grid infrastructure and interconnecting mains reinforcing; base periods decrease in the energy market; improvement of methods for forecasting wind speed and solar insulation energy flux density, as well as increasing the forecasting meteorological information arrival rate [9].

The choice of a WPP/ SPP generation forecasting method is carried out according to the forecasting time-frame. In [10], the author provides an in-depth review of existing approaches to predicting the solar insulation energy flux density, and also gives instructions on their use, depending on the required forecasting time-frame:

- up to 60 minutes - a combination of a physical approach (data from satellite and cloud cameras) with models based on regression analysis;
- up to 24 hours - methods of processing statistical data;
- from 24 hours - models and methods of numerical weather forecasting.

Regression analysis methods, as a rule, run well both in data-constrained environment and their redundancy [11]. In data-constrained environment, only historical meteorological data or power plant capacity can be accessed, which reduces the generation forecast accuracy. Using forecasting meteorological information from hydrometeorological centers, it is possible to increase the SPP generation forecasting accuracy and, as a result, to increase the power system efficiency. [12]

### B. Current situation in Russia

Russia is not the world-first in RES integration into the power system, but nevertheless, the government of the Russian Federation (RF) realizes the importance of

diversifying energy supply sources for the country's energy security and in [13] formulates targets for 2024. According to [13], RES facilities should produce 4.5% of electricity from the total generation of the UPS of Russia, and the total installed capacity should be 5,342 MW.

To achieve these goals, the Government of the RF provides incentive measures to support the generating facilities based on RES development and use.

- Wholesale energy market

Obligatory for buyers purchase volume of electricity from RES generating facilities and an addition to the wholesale market equilibrium price.

- Retail energy market

Priority electricity purchase by grid organizations from RES generation facilities at tariffs to compensate for losses in power grids.

- New generating capacities commissioning

Long-term power delivery contract (RES LPDC).

- Utility connection

Subsidies from the federal budget to compensate for utility connection for RES generating facilities with an installed capacity of no more than 25 MW.

- Local manufacturing

Target indicators of main and auxiliary equipment local manufacturing until 2024 have been determined, which will reduce the construction cost and increase the RES generation competition. A map of Russia with RES generation facilities (over 10 MW) is shown in Fig. 1-2.



Fig. 1. SPP location on the territory of the Russian Federation (in yellow - active, in grey - under construction)

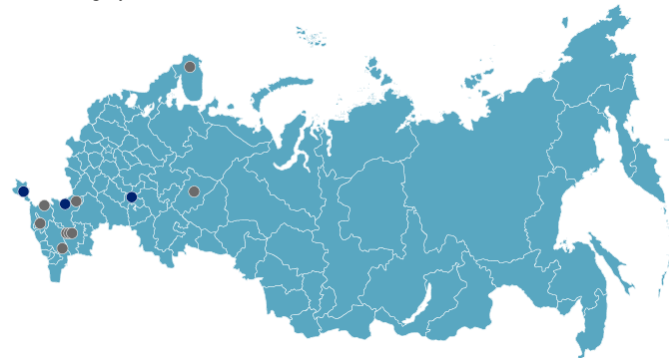


Fig. 2. WPP location on the territory of the Russian Federation (in yellow - active, in grey - under construction)

It can be seen from the figures above that today SPP are mainly used. The SPP/WPP ratio in the Integrated Power Systems (IPS) of the Russian Federation today is shown in Fig. 3.

Considering that the southern regions (IPS of the South) are the best option for the SPP localization, it is obvious that in these territories their number will increase over time. As a result, it is necessary to monitor the SPP operation impact on the power system and develop optimal methods for their accounting in the short-term power planning, even if at present SPP does not have a significant impact.

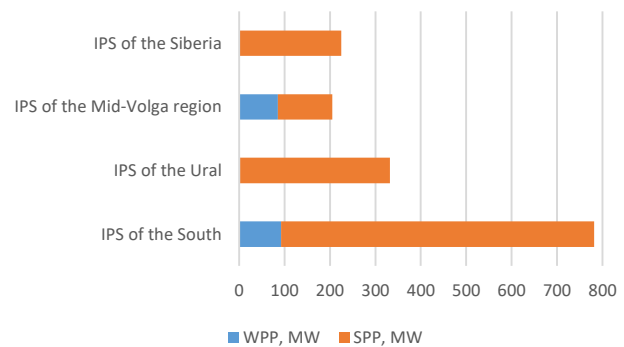


Fig. 3. The SPP/WPP ratio in the Integrated Power Systems (IPS) of the RF, MW

Due to the fact that the RES facilities commissioning in the Russian Federation is often not accompanied by meteorological stations (meteorological desks) multiplication directly at the RES location, an important task is to determine the SPP generation forecasting using only the available meteorological data from the system operator. This forecast can also be used to change in real time the maximum allowed power flow value (APFV) in controlled sections (CS), which was formed using RES facilities in the calculations.

Such a mechanism will optimize the IPS functioning and reduce the price indices of the balancing energy market. [10]

### III. CASE-STUDY

As a study object, a real power system X was selected, with the installed capacity of 3915 MW, 330 MW of which is the solar power plants share, and is shown in Fig. 4. On the operational zone territory there are four meteorological stations, which record and transmit forecast information to the system operator in the amount as presented in Table. I.

TABLE I. METEOROLOGICAL DATA FOR SPP GENERATION FORECAST

No	Parameter	Periodicity
1.	Temperature	1 hour
2.	Cloudiness	3 hours
3.	Weather elements	3 hours

The system operator uses the SPP generation forecast on a 10-minute lead-in interval to analyze the active power decrease, to take it into account when calculating the static aperiodic stability as a possible regular disturbance in the power system.

It is quite difficult to achieve a high accuracy of the SPP generation forecasting on a 10-minute lead-in interval on the data of Table I with their periodicity of 1-3 hours. However, this is a real and necessary task to determine the dependence

of power on changes in weather conditions and the influence of each weather parameter on the SPP power output.

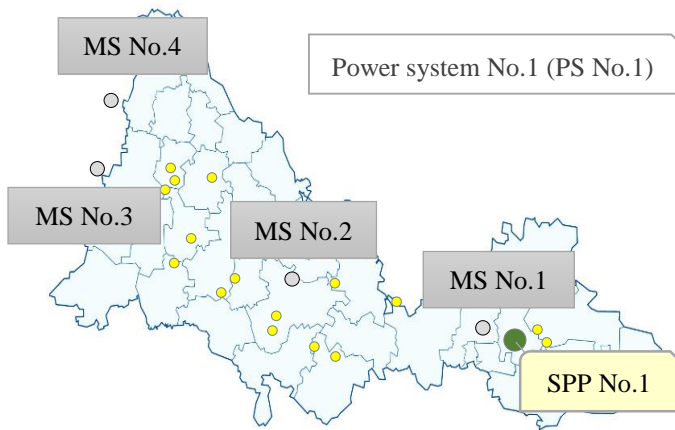


Fig. 4. Meteorological stations location from which the system operator receives information

#### A. Analysis of the change in the balancing market average price indice.

As mentioned above, the electricity price in the balancing market may depend on the accuracy and forecasting time-frame. To confirm this hypothesis, a comparison of nodal prices was carried out through the balancing market average price indices (BMI) of power plants located in power system No. 1 for the period from 2016 to 2020 in different regions of this power system. Within this study, they made an analysis of the average price change in areas with a high share of RES and in the areas without RES facilities. 9 in total different stations were analyzed.

Due to the fact that the SPP full potential can be manifested in the summer period, it was supposed to look into prices for April - September. In order to correctly take into account the changes in the BMI, including 2020, the prices for August - September were not taken into account.

As a result, it was revealed that all BMI variance graphs have an identical shape, similar to that shown in Fig. 5.

In order to get rid of the common for all power plants price change factors, the general form factors were obtained, which were subtracted from each graph. They use the mathematical tool to isolate the additive seasonal component ( $S$ ) from the additivity equation in accordance with expression (1):

$$F = T + S, \quad (1)$$

$F$  – the SPP generation forecast;  $T$  – the trend;  $S$  – the seasonality.

Average price graphs excluding the seasonal component are shown in Fig. 6, and the relative coefficients of price change for 2016 are presented in Table. 2.

It can be seen from Fig. 6 that the trends in the change in the average price for power plants No. 1 and from other power plants have similar slope coefficients. This means that there is no RES facilities influence on the balancing market price at their current capacities in the energy system.

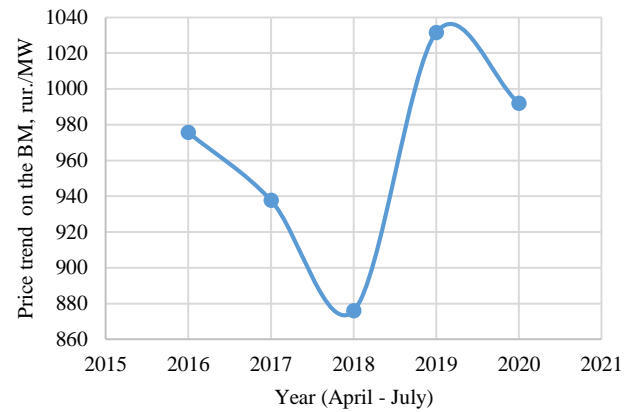


Fig. 5. Standard BMI variance graphs

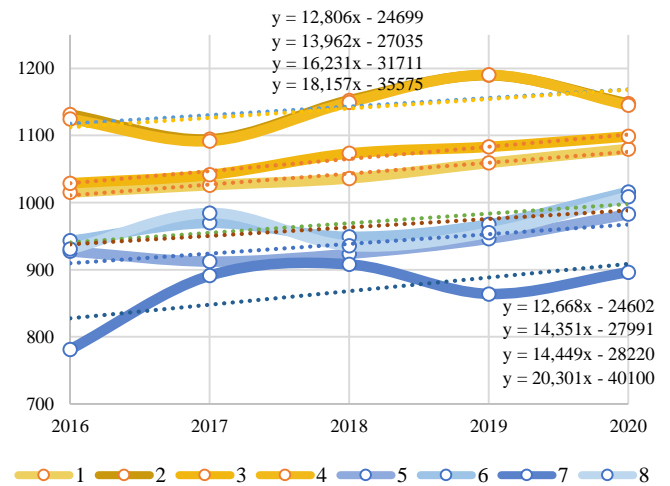


Fig. 6. The average price index in the balancing market

TABLE II. BMI CHANGES FOR 2016

Power plant's No.	PS No.	2016	2017	2018	2019	2020
1	1	1	1,01	1,02	1,03	1,05
2	1	1	1,01	1,03	1,04	1,05
3	1	1	1,02	1,03	1,05	1,06
4	1	1	1,02	1,04	1,05	1,07
5	2	1	1,01	1,03	1,04	1,05
6	3	1	1,02	1,03	1,05	1,06
7	4	1	1,02	1,03	1,05	1,06
8	5	1	1,02	1,05	1,07	1,10

#### B. Analysis of the forecast parameters influence on the SPP generation

The need to form a forecasting model is caused by the requirements of methodological guidelines on the power systems stability to take into account the change in RES facilities power output at an interval of 10 minutes [15].

The analysis of the forecasting parameters influence on the SPP generation was carried out through the built-in tool in Excel. Preliminary, a typical dome-shaped SPP generation schedule was divided into two parts (Fig. 7): ascending (growth) and descending (decline). Regression analysis was applied for each part of the graph.



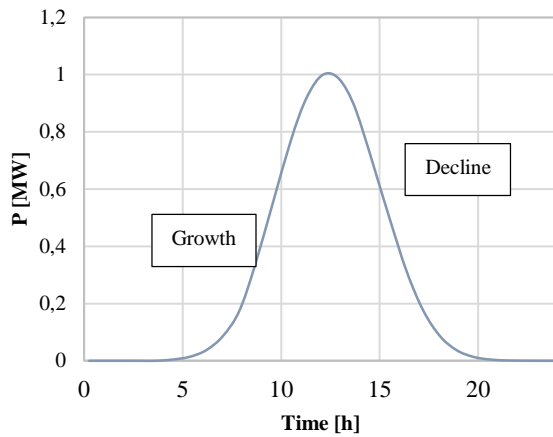


Fig. 7. Typical SPP power output schedule

Regression analysis shows the effect of some values (independent) on the dependent variable. The linear regression model is as follows:

$$P = a \cdot X_1 + b \cdot X_2 + c \cdot X_3 + d, \quad (2)$$

where  $X_1, X_2, X_3$  – the model parameters (temperature, cloudiness, weather elements);  $a, b, c, d$  – the model coefficients.

The data analysis function results in the following coefficients:

- $R$ -squared – determination coefficient. The higher the determination coefficient is, the better the model is.
- Coefficient  $D$  shows what  $P$  will be if all variables in the model are equal to 0. That is, other factors that are not described in the model also affect the value of the analyzed parameter.
- Coefficient  $A$  shows how much the parameter influences  $P$ .

As a study result, graphs of the linear regression model parameters changes in the context of one year were obtained (Fig. 8-11). Analysis and interpretation of the results will make it possible to talk about the quality of the model and the possibility of its application in calculating electric power modes.

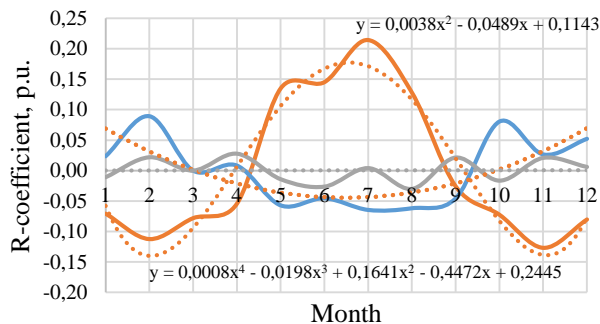


Fig 8.  $R$  coefficient for the ascending part of the graph: orange- the temperature; blue- the cloudiness; grey- the weather elements; orange dashed- polynomial characteristic(temperature); yellow dashed- polynomial characteristic (cloudiness); grey dashed- linear characteristic( weather elements).

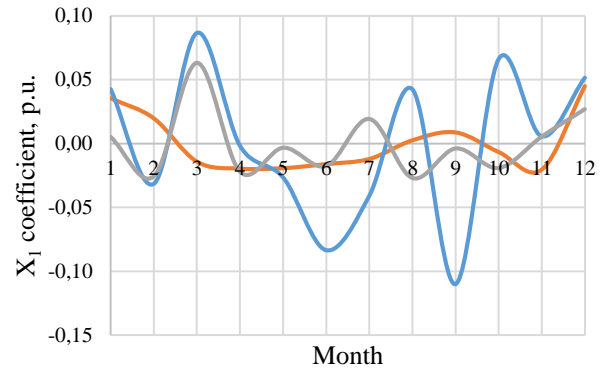


Fig 9.  $X_1$  coefficient for the ascending part of the graph: orange- the temperature; blue- the cloudiness; grey- the weather elements.

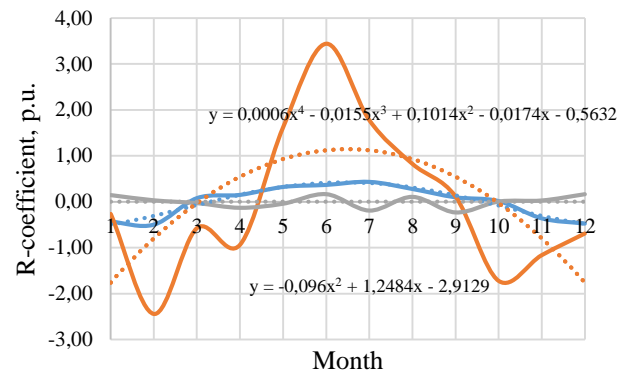


Fig 10.  $R$  coefficient for the descending part of the graph: orange- the temperature; blue- the cloudiness; grey- the weather elements; orange dashed- polynomial characteristic(temperature); yellow dashed- polynomial characteristic (cloudiness); grey dashed- linear characteristic( weather elements);

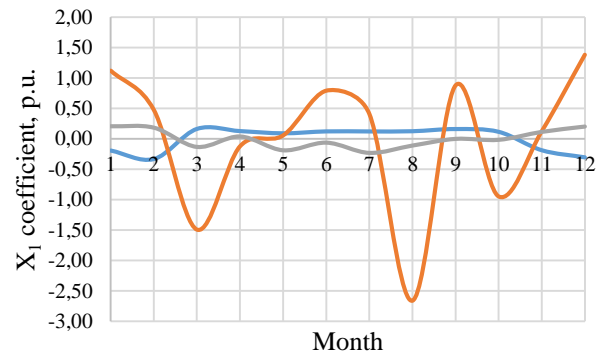


Fig 11.  $X_1$  coefficient for the descending part of the graph: orange- the temperature; blue- the cloudiness; grey- the weather elements.

The validity assessment can be given by analyzing the model coefficients as a whole. In the period from May to August, the coefficient  $Y$  ( $X=0$ ) of the “Temperature” parameter has negative values. That is, in the absence of this parameter in the model, the SPP power will have negative values. Since this is impossible, the values of all coefficients for these months of the “Temperature” parameter can be considered invalid.

The coefficient  $X_1$  characterizes the relation between the SPP power output and the model parameter. According to Fig. 8-9 such parameters as “Cloudiness” and “Weather elements”

have a feed-back with the SPP power output, and “Temperature” - a feed-forward. It should also be noted that for the descending part of the graph, the influence of cloudiness is more significant than for the ascending one.

The  $R$  coefficient of the model parameters for both parts of the graph - ascending and descending (Fig. 10-11), have similar values and behavior. It can be seen that the greatest influence on the model quality is exerted by “Clouds” with a greater influence in the winter than in the summer.

Based on the results obtained, we can conclude that the “Cloudiness” parameter has the greatest weighting on the SPP power output change and the effect on the model quality. Thus, the system operator needs to increase the forecasting metainformation arrival to improve the SPP generation forecasting model, which can be used in calculating electric power modes.

#### IV. CONCLUSIONS

The available meteorological data (temperature, cloudiness, weather elements) used by the system operator are poorly formalized into a mathematical model. Cloudiness has the greatest influence on the SPP capacity change. Cloudiness forecast has the highest value of the coefficient  $X_1$ , which characterizes the influence of other parameters.

Considering that the southern regions of Russia are the best locations for installing solar panels, the RES facilities localization in a certain area will increase. It appears that it is necessary to monitor the SPP influence on the power system and develop optimal methods for their accounting in the short-term power planning, even if the current availability of SPP does not have a significant impact.

The forecasting meteorological information arrival interval (3 hours) is too long for adequate SPP generation forecasting. It is necessary to reduce the interval of information arrival (5-30 minutes).

#### REFERENCES

- [1] J. Dudiak, M. Kolcun, “Integration of renewable energy sources to the power system,” in *EEEIC*, Krakow, Poland, May 10-12, 2014.
- [2] C. Breuer, C. Engelhardt and A. Moser, “Expectation-based reserve capacity dimensioning in power systems with an increasing intermittent feed-in,” in *Proc. European Energy Markets*, Stockholm/Sweden, May 27–31, 2013.
- [3] A. Navon, P. Kulbekov, S. Dolev, G. Yehuda, Y. Levron. (2020, May). Integration of distributed renewable energy sources in Israel: Transmission congestion challenges and policy recommendations. *Energy Policy*, Elsevier. [Online]. Volume 140. Available: <https://doi.org/10.1016/j.enpol.2020.111412>.
- [4] M. Abuella, B. Chowdhury, “Solar power probabilistic forecasting by using multiple linear regression analysis,” in *Proc. IEEE SoutheastCon 2015*, Fort Lauderdale/USA, Apr. 9–12, 2015.
- [5] J. Kleissl, *Solar Energy Forecasting and Resource Assessment*. Elsevier, 2013.
- [6] M. Milligan, M. N. Schwartz, and Y. Wan, “Statistical wind power forecasting for U.S. wind farms,” in *Proc. 17th Conf. Probability Stat. Atmos. Sci.*, Seattle, WA, Jan. 2004.
- [7] M. K. Behera, I. Majumder, and N. Nayak, “Solar photovoltaic power forecasting using optimized modified extreme learning machine technique,” *Eng. Sci. Technol. an Int. J.*, vol. 21, no. 3, pp. 428–438, Jun. 2018.
- [8] Z. Wen, Y. Li, Y. Tan, Y. Cao, “A combined forecasting method for renewable generations and loads in power system,” in *APPEEC*, Brisbane, QLD, Australia, Nov. 15-18, 2015.
- [9] Gerard Wynn. “Power-industry. Transition, here and now. Wind and solar won’t break the grid: nine case studies”, 2018.
- [10] Diagne, M., M.David, P.Lauret, J.Boland, and N.Schmutz, 2013: Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable Sustainable Energy Rev.*, 27, 65–76, doi:<https://doi.org/10.1016/j.rser.2013.06.042>.
- [11] Inman R. H., H. T. C. Pedro, and C. F. M. Coimbra, 2013: Solar forecasting methods for renewable energy integration. *Prog. Energy Combust. Sci.*, 39, 535–576, doi:<https://doi.org/10.1016/j.pecs.2013.06.002>.
- [12] T. Schröter, A. Richter and M. Wolter, “Development of Methods for an Optimized Infeed Forecast of Renewable Energies,” in *PMAPS*, Boise, ID, USA, June 24-28, 2018. 11
- [13] On approval of the Major public policy in improving the energy efficiency of the electric power industry based on renewable energy sources for the period until 2024, 2019.
- [14] D. R. Graeber, “Handel mit Strom aus erneuerbaren Energien”, Wiesbaden, Springer Fachmedien Wiesbaden, 2014.
- [15] Guidelines for the power systems sustainability, 2018.

# Vehicle Body Design and Analysis Aerodynamic by Flow Simulation

Phu Thuong Luu Nguyen

Automotive Engineering Department

Ho Chi Minh City University of Technology (HUTECH)

Ho Chi Minh, Vietnam

npt.luu@hutech.edu.vn

Van Dung Do

Automotive Engineering Department

Ho Chi Minh City University of Technology and Education

Ho Chi Minh, Vietnam

dodzung@hcmute.edu.vn

**Abstract**—In this research presents the vehicle body surface design and analysis of vehicle aerodynamic using flow simulation. The purpose of this paper is to develop a method for surface design of vehicle body model in SOLIDWORK software by using 2D real car images. Then this body structure is utilized for simulation and aerodynamic analysis using flow air. In this article, flow simulation/SOLIDWORK was used to analyze and to make the vehicle model to achieve more realistic simulation results. The model used in the simulation of this air flow has the dimensions as well as the mass, materials and properties that are almost equivalent to real life to evaluate the results of the air currents acting on the vehicle, at each other level of speed from 40km/h to 120 km/h. This paper is intended to provide some experience and familiarity with flow analysis and simulation.

**Keywords**—Flow simulation, Analysis, Vehicle, Design, Dynamics

## I. INTRODUCTION

Due to the extremely competitive nature of the automotive engineering industry, leading to the strong promotion of the technical and service sectors of each car manufacturer in the world, requiring the production of safety cars. At the present the car manufactures are more comprehensive, competitively priced, higher operating efficiency in a shorter time period while at the same time low maintenance costs. Especially with a big car manufacturer like Mercedes-Benz, it motivates its leading engineers to learn and make the most of the durable properties of the materials, the aerodynamic methods, to adapt to the customer requirements.

This also necessitates automotive engineers to consistently investigate safety and reliability characteristics of the vehicle to reach the optimum designs which can meet often conflicting design criteria. In recent years, use of computer-aided engineering (CAE) methods has been a good asset for engineers, particularly, at the early stages of body-in-white (BIW) structure design. Several literatures have been contributed to proposing innovative approaches as well as models for optimal simplification of vehicle body design [1-5].

Nowadays, with the great development of science and technology, the help of computer technology helps in calculating simulation algorithms. Many technical documents and documents have values for reference; many solutions have been applied in practice to optimize the bodywork, due to exposure to air pressure and extreme weather conditions. In this study, with the help of solid work

software, the Mercedes-Benz E200 2016 version will be simulated to analyze the aerodynamics at different speeds from 40km/h to 120km/h.

## II. BUILDING THE 3D VEHICLE BODY MODEL

### A. 2D Sketch

The vehicle body investigated in this study is a Sedan car. Those dimensions are shown on the Mercedes E 200 website of Mercedes company. The total length of the vehicle is  $L = 4818$  mm, the height is 1440 mm and wide is 1536 mm. The distance between front and rear wheel is 2833 mm. The geometrical configuration of the vehicle body 2D image is shown in Fig. 1.

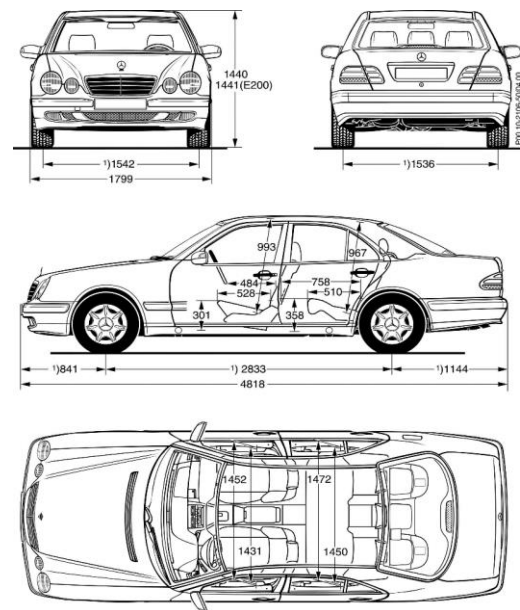


Fig. 1. The 2D image of vehicle body model.

### B. 3D sketch vehicle body model designs

The 2D image in Fig. 1 was used for the surface method design 3D model. This way we will be able to draw the car with the highest proportions and accuracy, use the Spline command and switch to 3D Sketch mode. After exporting the picture, we start to adjust the parameters of the following image to best suit the actual size of the car in real life. Pay particular attention to the composition of the aerial view, so that the center of gravity is in the center of the vehicle as shown in Fig. 2.

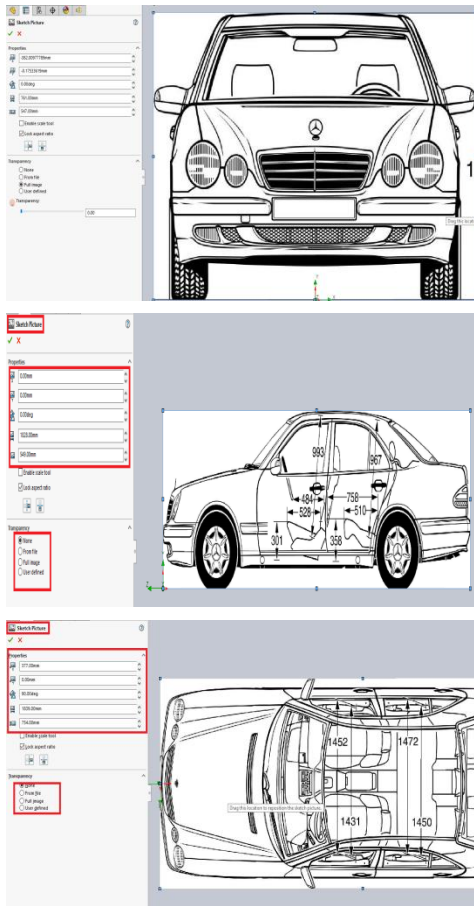


Fig. 2. The 2D image Front, Right and Top of vehicle model in sketch picture.

After creating 3 drawings on 3 planes, we continue to create a symmetry with the front view to draw the rear of the vehicle, use the Reference Geometry→Plane command and then perform the following steps: 3D vehicle model as shown in Fig 3, Just draw 3DSketch using Spline command on 1/2 of the vehicle, then use Linear Pattern → Mirror → Mirror Plane (right plane) → Body to Mirror command and select the parts you want to Mirror as shown in Fig. 4.

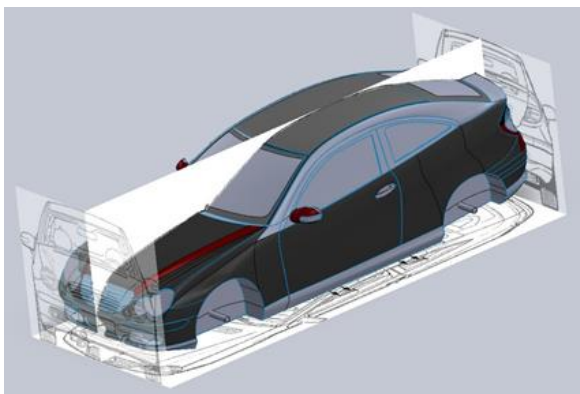


Fig. 3. The surface method for 3D model design

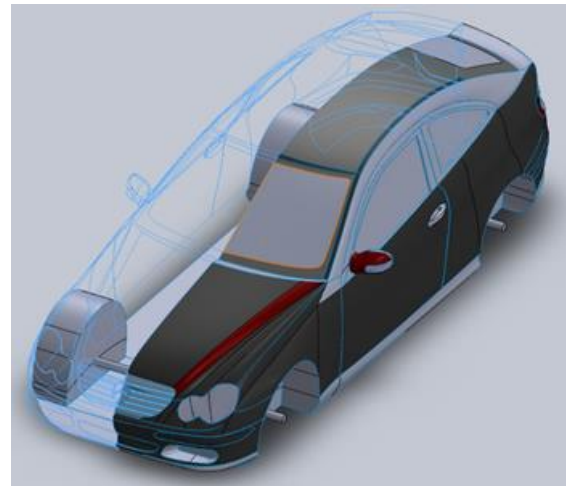


Fig. 4. The half of 3D model design

### C. Apply materials in wizard

We can apply the material in the model by following the steps as shown in Figures below.

In Fig. 5, we will name the simulation, choose the format and always set: Use Current and default. In Fig 6, here we will choose the calculation unit and this will affect during the calculation as well as the results of the simulation process so we choose the SI unit (m-k-g-s) and click Next. Fig. 7: The setting of external, at here, we will select the effect up to car and are shared doing 2 types: internal and external and selected External --considers close cavities-- selected exclude cavities and exclude internal space. Fig. 8: Adding Air for Flow simulation, Here, we will choose the form of the impact force and specifically the air when the car moves at high speed, select Gases  $\rightarrow$  Air  $\rightarrow$  Add and then click next to proceed with the direction of the air flow. In addition, we need to pay attention to select the Laminar and turbulent, these are the basic properties of Air. Fig. 9: Setting wall condition, we will add more properties to the simulated wall and since this is an air flow simulation, we leave the default mode including: value (parameter), Adiabatic wall (Default wall thermal condition), Roughness (0 micrometer).

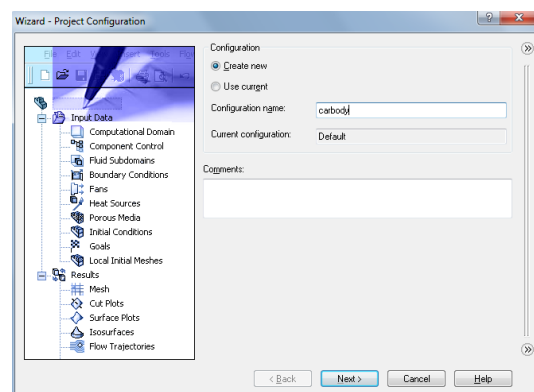


Fig. 5. The applied properties of project configuration

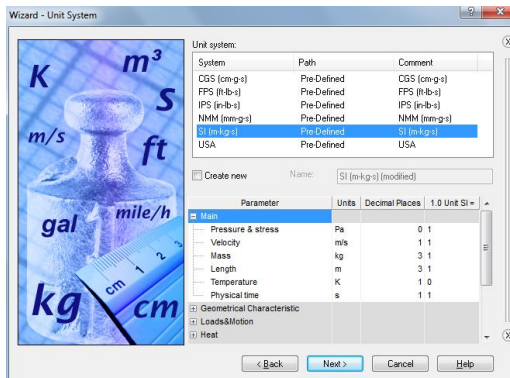


Fig. 6. The chosen unit system

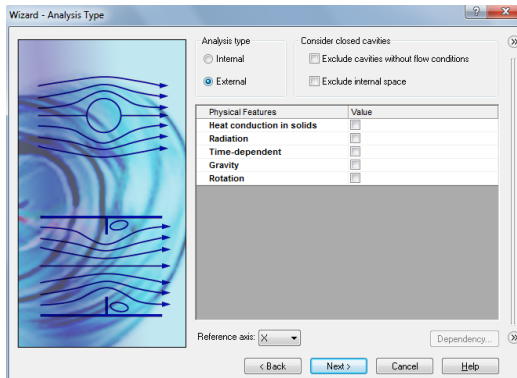


Fig. 7. The setting of external analysis

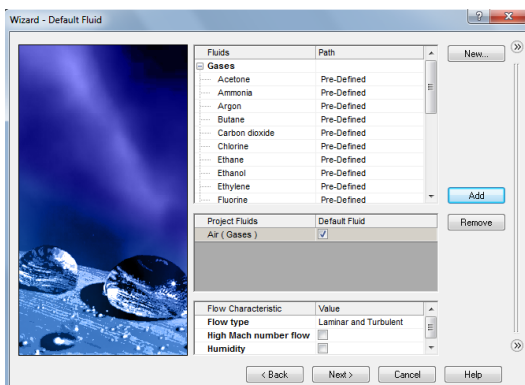


Fig. 8. Fig 8: Adding the air for simulation

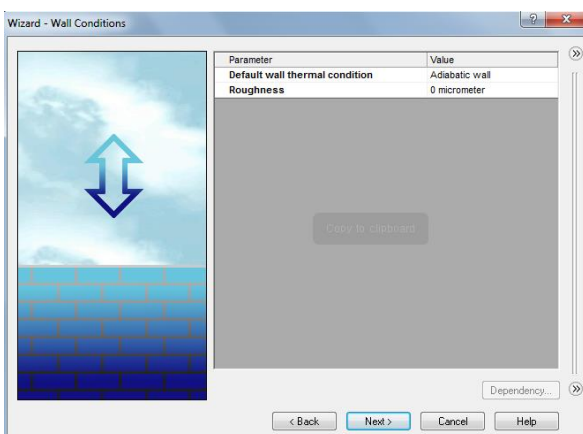


Fig. 9. Setting wall conditions

Fig. 10: Setting the initial and ambient conditions with velocity. First, pay attention to the direction of the vehicle in Solid work, if the direction of the vehicle follows the Z axis, go to Coordinate System → Global Coordinate System and

the Reference axis choose Z. Because the air movement will be in the opposite direction of the vehicle's movement or the value of the velocity on the Z axis will always be negative, at the same time start to choose values of speed 40Km/h, 60Km/h, 80Km/h, respectively. Then setting the results and geometry resolution as shown in Fig. 11.

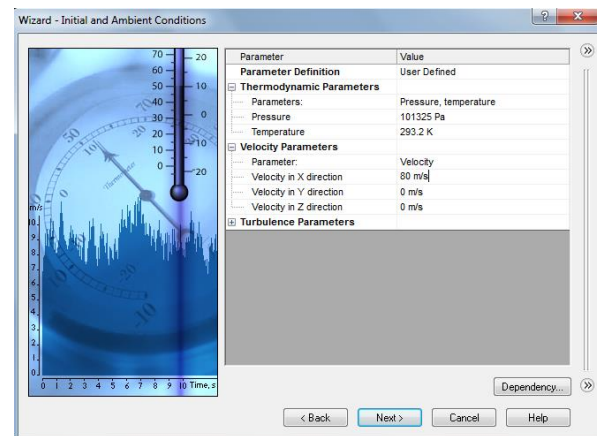


Fig. 10. Setting the initial and ambient conditions with velocity 80 m/s

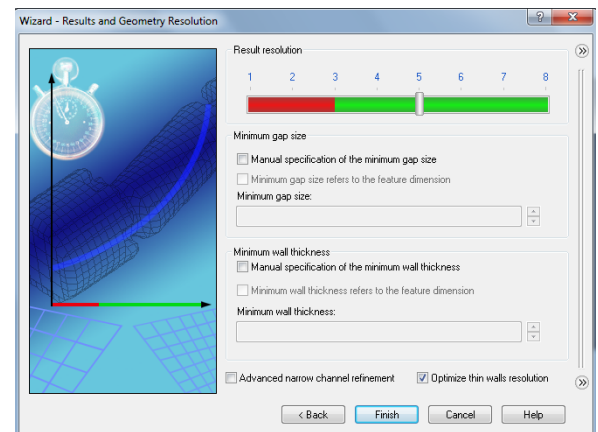


Fig. 11. Setting the results and geometry resolution

#### D. Finite element model computational domain

In this part the computational domain was presented as shown in Fig. 12. This helps to fix the simulation space while optimizing the simulation capabilities. That is a simulation space, consisting of three X Y Z axes and divided into negative and positive, the purpose of optimizing the simulation space so that the air flow does not move chaotically.

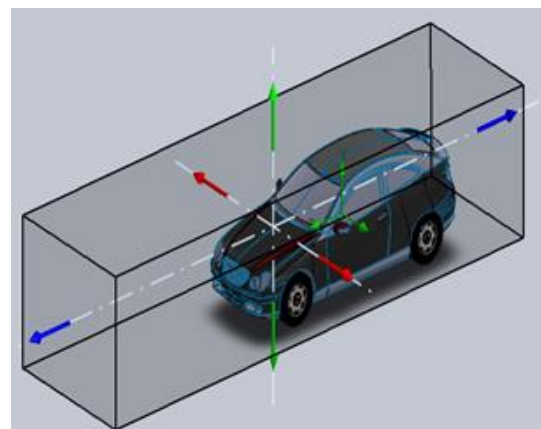


Fig. 12. The adjusting computational domain



There will be Size and conditional to adjust the size of the space including the X, -X axes; Y, -Y; Z, -Z. where X, -X are symmetrical Y, -Y is symmetrical, but Z, -Z, the length of Z is greater than -Z because this is the direction of the gas flow as shown in Fig. 13. We will apply the forces to the vehicle, including the force acting on the XYZ axes, in the insert goal parameter there are many types of force, but we just need to focus on Normal Force (X), (Y) and (Z). Choose the three types of force as shown in Fig. 14.

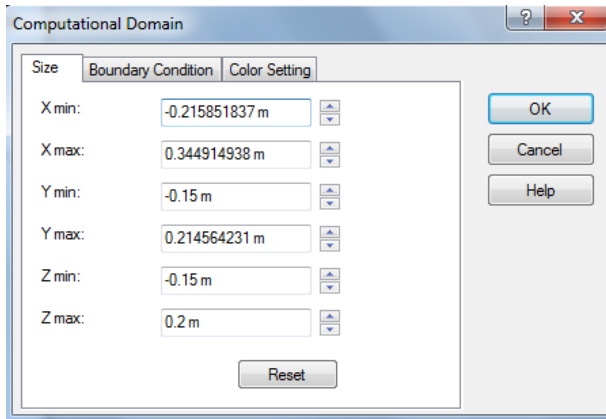


Fig. 13. The computational domain size

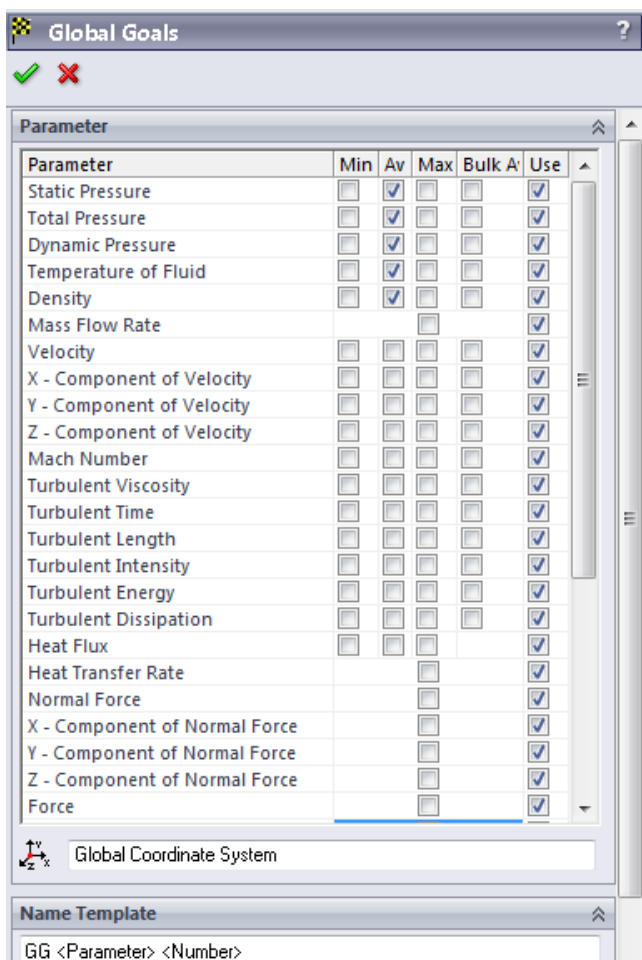


Fig. 14. Insert goal parameters

#### E. Set up interactions and run

In this section, this research presented how to set the interactions and run the project. The interaction time depends on the smooth simulation. In this paper set 10 for interaction

because of the calculation time as shown in Fig 15. After setting all the parameters and data as well as selecting the force base and direction, we simulate the vehicle. Press **RUN**, there will be a table to setup through the calculator. Select: Solve → New calculation → Run at: This computer → Use: All cores → RUN as shown in Fig. 16.

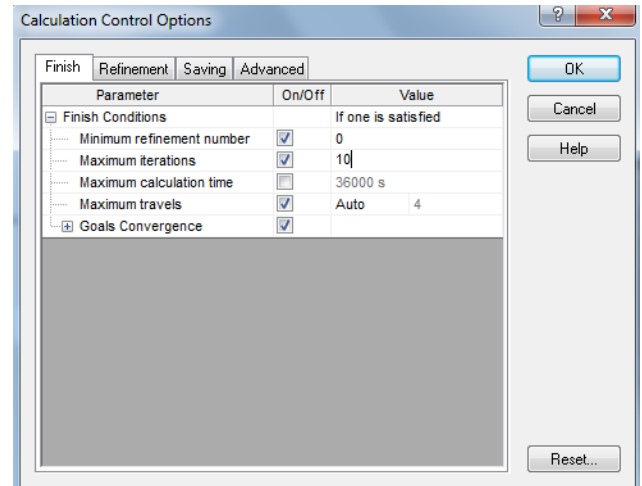


Fig. 15. The calculation control setting

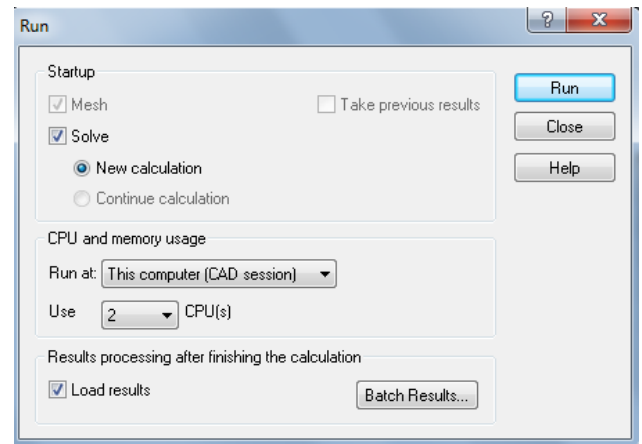


Fig. 16. The CPU setting for calculation

### III. CALCULATING AND SIMULATING RESULTS

The solving information was shown in Fig. 17. In this table, all solved results were presented such as warning errors, calculation time, running date. Information boards in each vehicle will be different, but the time is fast or slow depending on the Total cell, usually fluctuating in the range of 10-50 thousand and takes 2 minutes to 5 minutes depending on the number of cores used. After running the data on the computer, we will start the simulation. First, we will advance the cut Plot simulation, this simulation to simulate the acting velocity along the length of the vehicle, we choose the direction of the air movement along the vehicle and coincide with the center of gravity of the vehicle. After having vertical velocity, now go to the simulation of Surface Plot, the aim is to determine the magnitude of the drag on the vehicle when traveling at high speed, the distribution of force on the surface. Vehicle is used surface design as shown in Fig. 18.

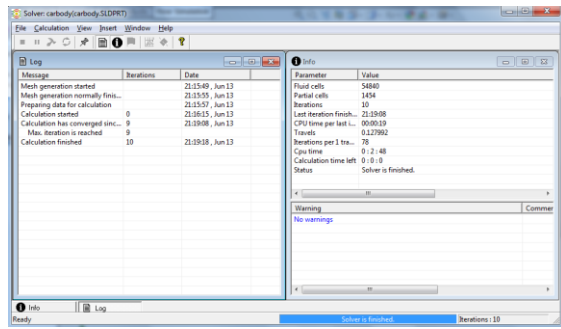


Fig. 17. The calculation solving time

To see the simulation we have to set the flow trajectories as shown in Fig 18.

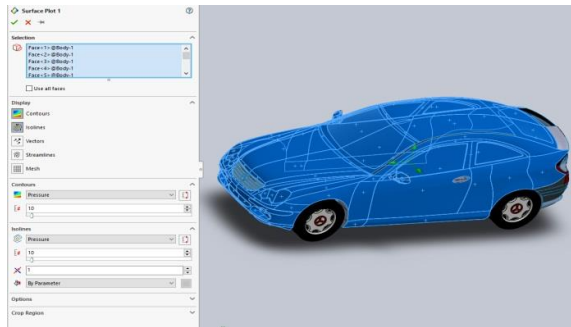


Fig. 18. The flow trajectories setup parameters

The results of pressure and temperature simulation are shown in figure from 19 to 23 of this section. To see different results, we just change the setting of velocity from 40 km/h to 120 km/h.

And this is a typical example of force exerted on a vehicle; it manifests itself in a range of colors from smallest to largest. Cut Plot's impact is evenly spread across the vehicle, but always has the maximum and minimum values shown in the Fig. 19, the color range from blue to red. Finally, the Flow Trajectories, with this simulation, will simulate the wind speed in the form of long arrow lines evenly distributed across the vehicle's surface from the Front View direction. The purpose of this is to simulate the flow of air impacting the vehicle and check the perfection of the vehicle's parts. If the arrow beams directed at the vehicle are dispersed when it touches the surface of the vehicle, passed, otherwise failed and the number of arrows commonly used in this simulation will be in the range 100-150 as shown in Fig. 20.

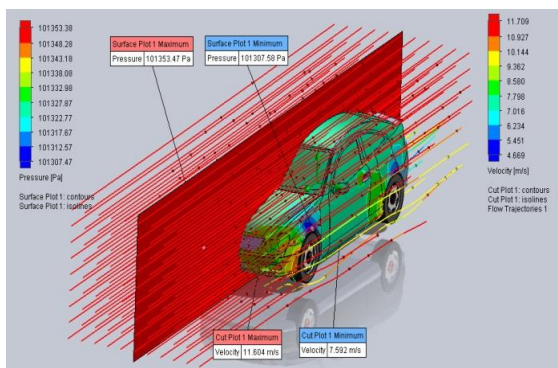


Fig. 19. The pressure, velocity and flow trajectories surface plot contour at 40 km/h

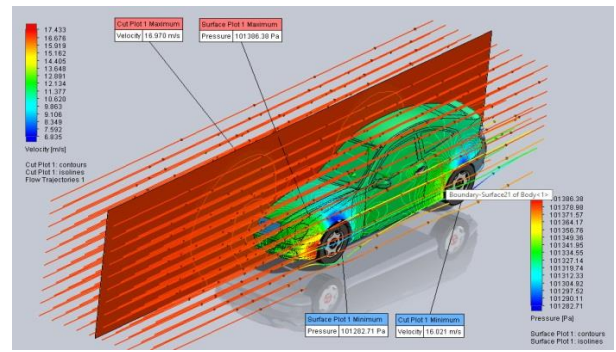


Fig. 20. The pressure, velocity and flow trajectories surface plot contour at 60 km/h

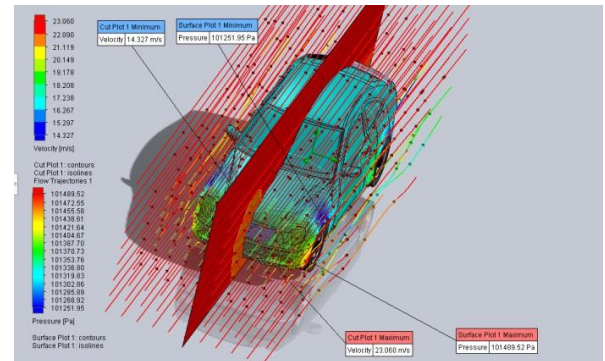


Fig. 21. The pressure, velocity and flow trajectories surface plot contour at 80 km/h

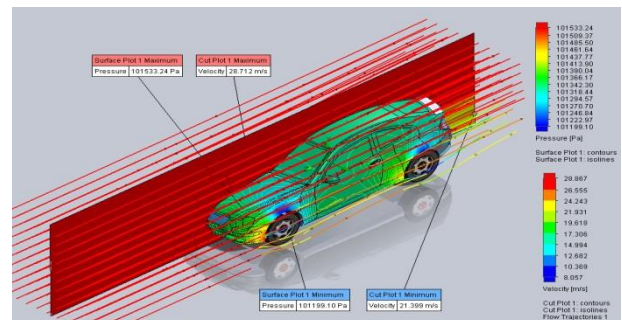


Fig. 22. The pressure, velocity and flow trajectories surface plot contour at 100 km/h

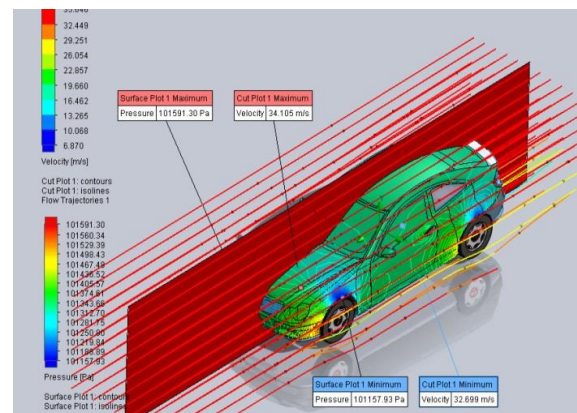


Fig. 23. The pressure, velocity and flow trajectories surface plot contour at 120 km/h

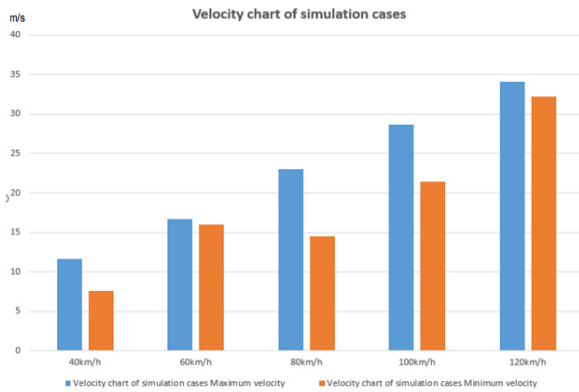


Fig. 24. The velocity chart of simulation with different cases

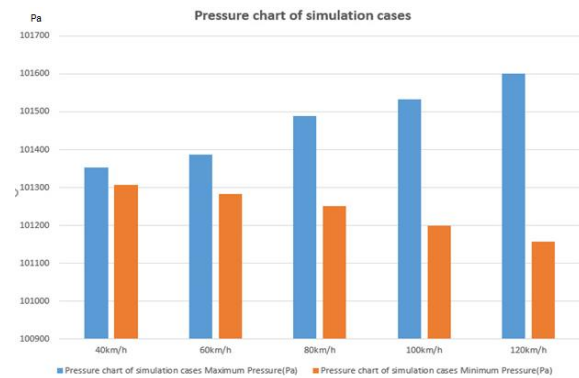


Fig. 25. The pressure chart of simulation with different cases

The Surface Plot, which simulates the force of velocity exerted on the surface of the vehicle. We choose as much force contact surface as possible. Similar to cut Plot, to get the most accurate values we need to Reset Surface to maximum and Reset Surface to Minimum.

In this paper, the different speeds of air flow were performed from 40 km/h to 120 km/h and the values of each case as shown in table 1 and Fig 24 to Fig. 25. When we make the simulated velocity for each case, it can be seen that the maximum velocity of each case never exceeds the rated speed is always lower than 5-10m/s. The vehicle has a rating mechanism to avoid overloading and wasting fuel so capacity caused by the conversion of mechanical energy into heat.

Besides, it can be seen that the minimum speed always has a difference with the maximum speed of about 40% (except for case 2) due to the effect of air and road surface resistance, and also the drag of engine details inside the vehicle.

Due to the structure of modern vehicles, there are wavy lines stretching on the body cap, helping to avoid passing the car at least, reducing air friction, so the maximum speed is always high.

The results show the difference between the speed and pressure on the vehicle. If the speed is proportional, the maximum speed and the minimum speed always increase more or less. Then, the pressure is inversely proportional. Fig. 24 and Fig. 25 show that, between the maximum pressures is always increasing case by case and the minimum pressure is always decreasing from the box magnetic field.

In the first case, the minimum pressure is reduced by 10% compared to the maximum pressure, then in case 5 120km / h, the minimum pressure is reduced by 60% compared to the maximum pressure.

This happens because when traveling at high speed, with the special construction of the vehicle with many wavy lines, it helps to concentrate the pressure on the front of the vehicle, reducing the pressure dispersed around.

TABLE I. THE COMPARISON OF DIFFERENT SPEEDS

Case	Velocity	Pressure	Air Flow
40 km/h	Maximum cut plot: 11.604m/s Minimum cut plot: 7.592m/s	Maximum surface plot: 101353.47Pa Minimum surface plot: 101307.58Pa	at 40km/h, at 100 Line with Arrows moderately moving air currents in yellow and orange are between 9 and 10m/s. As shown in Fig 19.
60 km/h	Maximum cut plot: 16.7m/s Minimum cut plot: 16.012m/s	Maximum surface plot: 101386.38Pa Minimum surface plot: 101282.71Pa	60km/h equivalent to 16.7m/s, with 100 line with arrows obtained; blue: 9.16m/s (wheels), green: 12,134m/s (frame), yellow and orange: 14,405m/s and 15,152m/s (sides). As shown in Fig 20.
80 km/h	Maximum cut plot: 23.06m/s Minimum cut plot: 14.5m/s	Maximum surface plot: 101489.81Pa Minimum surface plot: 101251.95Pa	speed 80km/h equivalent to 22.3m/s, with 100 line with arrows to get results; blue: 16.27m/s (wheels and frame), green: 19.17m/s (behind the car and sides), red orange: 22.9m/s and 23.5m/s in front of the vehicle, ceiling and side mirrors. As shown in Fig 21.
100 km/h	Maximum cut plot: 28.7m/s Minimum cut plot: 21.4m/s	Maximum surface plot: 101533.24Pa Minimum surface plot: 101199 Pa	100km/h equivalent to 27m/s, with 100 line with arrows that will elastic simulation results; yellow: 24.24m/s (wheels), orange: 23-26.55m/s (both sides of the car) and red: 26.55-28.867m/s on the rest of the vehicle surface. As shown in Fig 22.
120 km/h	Maximum cut plot: 34.1m/s Minimum cut plot: 32.27m/s	Maximum surface plot: 101600Pa Minimum surface plot: 101157.93Pa	120km/h equivalent to 46m/s, with 100 line with arrows will achieve simulation results with 2 main colors; yellow 26.3-29.2m/s (both sides of the car) and red: 32.5-33.65m/s on the entire remaining surface of the vehicle. As shown in Fig 23.

#### IV. CONCLUDING REMARKS

In this research, the vehicle body surface design method was introduced and investigated. The models were formed based on different coupling of surface and beam elements as well as different extent of geometric simplification.

In the assessment, it will be divided into three categories for the Flow Simulation section including: Velocity, Pressure and Flow of the Flow Trajectories motion.

**Speed:** During the simulation, there is always an increase in speed from 40 to 120Km/h, through which we can see the impact of speed on the vehicle's operation. In the simulations the speed is divided by color from blue to red and the higher the speed, the percentage of the speedways passing the vehicle changes from blue to Green, orange and finally red. Especially the underbody and the front bumper where there are many angles. In the end, vehicle construction greatly affects wind and air resistance, and also the materials that make up the car.

**Pressure:** the higher the speed, the higher the pressure on the vehicle, but in general there is always a uniform pressure distribution on the E200, in all simulations, the car is always kept. The pressure is stable and no red streaks appear on the vehicle.

**Flow Trajectories,** which always go with the velocity of the air flow in the vehicle, the magnitude and strength of the

fluid flow is always proportional to the vehicle speed. At the same time when the vehicle is at high speed, the air currents will disperse to the sides of the vehicle to reduce friction with the durable air surface.

The implementation in SOLIDWORK was proved to simulation flow the air through the vehicle body, the ability to apply for optimal design vehicle body to minimum air drag force.

#### REFERENCES

- [1] N.P.T. Luu, ect., "Analysis of vehicle structural performance during small overlap frontal impact", IJAT, vol 16, pp. 799-805, 2015.
- [2] N.P.T. Luu, ect., "A study on optimal design of vehicles structure for improving small overlap rating", IJAT, vol 16, pp. 959-965, 2015.
- [3] Luu Nguyen Phu Thuong, "An optimisation approach to choose thickness of three members to improve IIHS small-overlap structural rating", IJ CRASHWORTHINESS, DOI: 10.1080/13588265.2017.1281203.
- [4] L. N. P. Thuong, "Vehicle Frontal Impact to Pole Barrier Simulation Using Computer Finite Element Model," 2018 4th International Conference on Green Technology and Sustainable Development (GTSD), Ho Chi Minh City, 2018, pp. 273-277, doi: 10.1109/GTSD.2018.8595702.
- [5] N. P. Thuong Luu, "Analysis of Bus Structural Performance During Full Frontal Impact," 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi, Vietnam, 2019, pp. 635-638, doi: 10.1109/ICSSE.2019.8823416.



# A Preliminary Study of a Two Stroke Free-Piston Engine for Electricity Generation

Nguyen Huynh Thi

Faculty of Vehicle and Energy  
Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
nguyenhuyhthi@tgu.edu.vn

Nguyen Van Trang

Faculty of Vehicle and Energy  
Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
trangnv@hcmute.edu.vn

Huynh Thanh Cong

Vietnam National University-Ho Chi  
Minh City  
Ho Chi Minh City, Vietnam  
htcong@hcmut.edu.vn

Huynh Van Loc

Faculty of Industrial Engineering,  
Tien Giang University  
My Tho City, Vietnam

Dao Huu Huy

Faculty of Vehicle and Energy  
Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam  
dhhuy2310@gmail.com

Ngo Duc Huy

Faculty of Vehicle and Energy  
Engineering,  
Ho Chi Minh City University of  
Technology and Education  
Ho Chi Minh City, Vietnam

**Abstract**— Vehicle contributes greatly to the rapid and strong economic development and other aspects of society. Besides, the operation of the transport impact on health and the environment seriously. Therefore, the study of solutions to improve fuel efficiency and environmental protection is one of the most focusing issues. The main objective of this paper is to present the theoretical fundamentals related to the design and development of a two-stroke free-piston engine (FPE) for electrical power generation. In this study, the prototype engine has been conducted with a power of 1.5kW, a bore size of 34mm, and a maximum possible stroke of 28mm using gasoline as fuel. In the future, this engine will be going to take advantage of excess biogas to generate electricity for production activities and daily life. The research has completed the initial intentions such as a theoretical investigation, calculation, determination of the engine model, and starting mechanism. Preliminary results have confirmed the feasibility of proposing the FPE model.

**Keywords**— *Free-piston engine, internal combustion engine, linear generator; starting system*

## I. INTRODUCTION

In recent years, along with the increase in population and the growth of economies, there has been a rapid increase in the number of private vehicles. When working, internal combustion engines emit many components that pollute the environment such as carbon dioxide (CO<sub>2</sub>), carbon monoxide (CO), nitric oxides (NO and NO<sub>2</sub>), sulfur dioxide (SO<sub>2</sub>), hydrocarbons and this substance in high concentrations will directly affect human health, the environment, and the atmosphere. This is one of the great challenges for the fields of internal combustion engines, energy, and the environment. Due to the limitations of fossil fuel and stricter emissions standards, conventional engines tend to replace by more efficient engines. FPE considers the solution with the advantages of reducing NO<sub>x</sub> emissions [1], multi-fuel [2], and high performance [3]. The absence of crankshaft makes the structure of FPE simple, reducing friction compared to conventional engines [4].

In recent years, many research groups are working on FPE prototypes as an alternative to conventional engines [5-

7, 19]. However, most prototypes are often complicated to control. To overcome these problems, a FPE prototype with a dual-piston configuration was proposed, by eliminating the generator, mechanical starting system, fixing top dead center (TDC) make it easy to manage some simple operating conditions.

## II. DEVELOPMENT OF ENGINE PROTOTYPE

FPE is the engine that converts from thermal energy to electric energy. FPE has variable compression ratio capability, thus it is suitable for multi-fuel operation [8]. Moreover, its high efficiency and quick transient response make it suitable for hybrid electric vehicle applications [9]. Currently, FPE configurations include single piston, dual piston, and opposed piston. However, the dual-piston configuration is capable of providing high performance despite being a relatively simple device [10]. This paper proposed a simple engine prototype of dual piston configuration in Fig. 1. The structure of the prototype includes two-chamber combustion, mechanical starting system, and driveshaft.

The engine works on the principle of the heat release of the first combustion chamber will create compression pressure in the second combustion chamber, then combustion chamber 2 will release heat by burning fuel with spark and continuous process. The main shaft brings the generator moved freely between two consecutive processes of heat release from two combustion chambers thereby generating electricity. In a dual piston configuration the heat release of the first piston greatly affects the second piston because the FPE is not restricted by the crankshaft connecting device, the piston dead center is not fixed. Therefore, without proper control, it can lead to a collision between the piston and cylinder head, and the unstable compression ratio can lead to deviations [11-12]. Therefore, mechanical mechanisms that manage the top dead center (TDC) and bottom dead center (BDC) was proposed to prevent the engine from losing control [18]. By eliminating the linear alternator, which will be replaced by a mechanical starting system that is easy to control.



The prototype uses two 2-stroke engines at both ends with a bore size of 34 mm and a piston stroke of 28 mm, the highest compression ratio is 7.5:1. The simple design of the free piston engine and the reduced number of moving parts minimize frictional losses as in Fig. 2. The crank mechanism will be eliminated, and the piston friction is reduced due to the purely linear motion, giving low side forces on the piston.

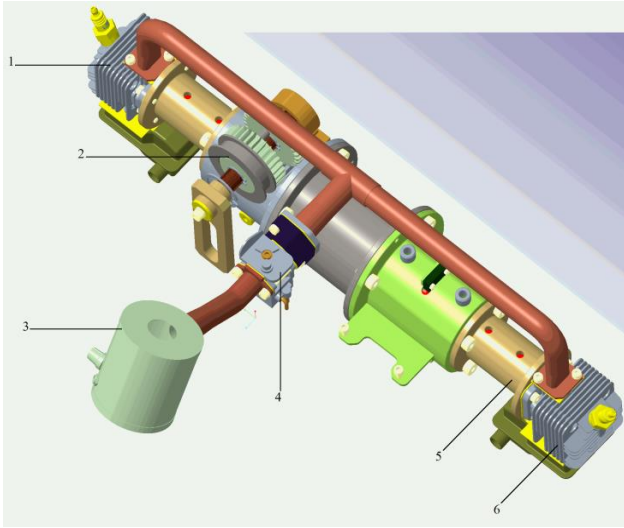


Fig. 1. A prototype of two-stroke free piston engine.

1- Cylinder 1; 2- Activation mechanism; 3- Compressor; 4-Carburetor; 5- Compressed air chamber; 6-Cylinder 2

A compressor is arranged to compress the air and fuel mixture into the compression chamber, the opening and closing of the discharge port and intake port will depend on the piston movement. The working cycle of the engine is composed of compression, combustion, expansion, and scavenging.

### III. CONCEPTUAL DESIGN OF FREE-PISTON ENGINE (FPE)

#### A. Thermodynamic cycle of FPE

The two-stroke engines work on the Otto cycle [13] which includes: two isochoric processes and two adiabatic processes. However, the correlation of the timing of every process differs from the four-stroke engine, and it depends on the cylinder's structure, which is the position of the transfer port and the exhaust port that showed in Fig. 2. The diameter of the cylinder is  $D$ ; the intake port diameter is  $D_i$  and the exhaust port diameter is  $D_e$ ;  $D_t$  is the transfer port size;  $S$  is the stroke length;  $S_t$  is the distance from the top dead center to the point starting open the transfer port;  $S_e$  is the distance from the top dead center to the point starting open the exhaust port;  $a$  is the point which the intake process begins;  $r$  is the compression process beginning point;  $c$  is the point which spark-plug ignites;  $z$  is the point which the cylinder's pressure reaches maximum  $p_z$ , and  $b$  is the beginning point of an exhausting process.

Assuming that the changing of the cylinder's pressure in the scavenging period is inconsiderable, it means that  $p_r \approx p_a$ . The thermodynamic calculation starts at the point  $r$  with the error of residual gas temperature  $T_r$  less than 1%.

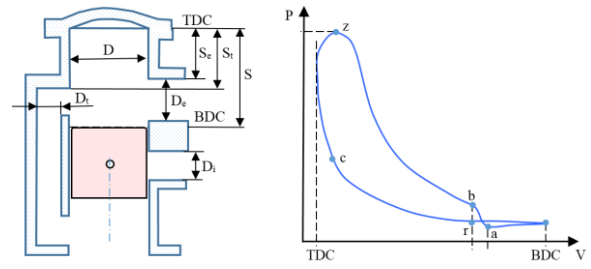


Fig. 2. Engine structure and the theoretical Otto cycle.

TABLE I. FREE-PISTON ENGINE SPECIFICATIONS

Number of cylinder	1
Type of engine	2 stroke
Bore	34 mm
Stroke	28 mm
Moving mass	0.58 kg
Nominal compression ratio	7.5:1
Fuel	gasoline

#### B. Simulate compressive pressure in the cylinder

Application in cylinders can be assessed by the derivative form of the first law of thermodynamics. The simulation will be based on Matlab/Simulink program.

$$\frac{dy}{dx} = -\gamma \frac{p}{V} \cdot \frac{dV}{dt} + (\gamma - 1) \frac{Q_{in}}{V} \frac{dx_b}{dt} \quad (1)$$

$P$  is pressure in cylinder (MPa);  $\gamma$  is the specific heat ratio;  $V$  is volume ( $m^3$ );  $Q_{in}$  is the input heat energy;  $x_b$  is mass of burned mixture.

The combustion process, which simulates the mass fraction burned, is performed by the Wiebe function [13-14,18].

$$x_b = 1 - \exp \left[ -a \cdot \left( \frac{t-t_s}{C_d} \right)^{b+1} \right] \quad (2)$$

$C_d$  is the combustion duration;  $t_s$  is ignition time. The constants of  $a = 5$ , and  $b = 2$ .

The simulation base on initial pressure taken from experiment 4.5 ( $kg/cm^2$ ); ignition time at 2 mm before TDC; Stoichiometric air – fuel ratio,  $\lambda \approx 1$ ; engine speed of 10 Hz. The results in Fig. 3 shows the relation between velocity and piston position displacement for mentioned compression ratios. The parameters change, but the curves keep the same shape. A larger compression ratio may lead to a bigger velocity and moving frequency.

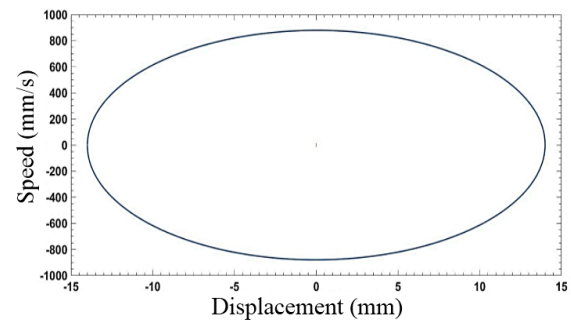


Fig. 3. The displacement corresponds to the velocity of the piston.

Simulation results can achieve pressure greater than 10  $kg/cm^2$  with the burn time ( $t-t_s$ ) 20 ms that can be shown in

Fig 4. Basing on the maximum pressure value, the ignition timing can be determined.

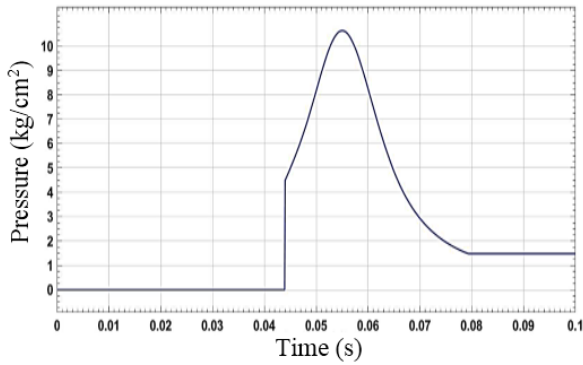


Fig. 4. Graph of peak pressure

#### IV. CONCEPTUAL DESIGN OF THE MECHANICAL STARTING SYSTEM

To startup, the FPE must be set to a certain frequency [15-17] to create pressure in the cylinder. Therefore, a mechanical starting system has been proposed with a simple control instead of a linear motor, to ensure piston displacement illustrated Fig. 5. The rack and pinion gears are used to convert rotational motion into reciprocating linear motion, a DC motor with a power of 500 W drive the driven gear to force the rack connected to the main shaft.

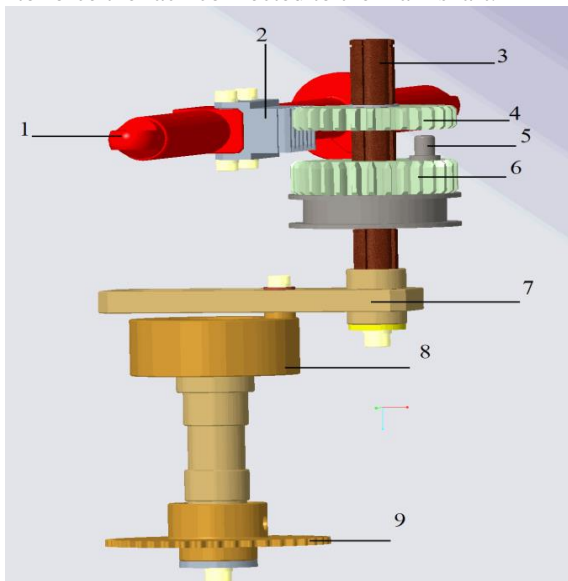


Fig 5. Mechanical starting mechanism.

1 - Main shaft; 2 - Rack; 3 - Shaft of starting mechanism, 4 - Driven gear; 5 - Fasteners; 6 - Drive gear; 7 - Handwheel; 8 - Eccentric wheel; 9 - Driven Gear from Starter Gear of Starter Motor.

The drive gears are linked to the rack and the main shaft when the main shaft moves the drive gear to determine the position of the main shaft. The driven gear is connected to the driven gear from the starter gear of the starter motor. When the driven gear and drive gear are interlocked through the fasteners, the force from the motor through the eccentric gear reaches the starting shaft and drives the main shaft. Moreover, the starting speed can be easily controlled by a motor controller that allows for starting of the engine at each different speed.

The rack pinion mechanism helps to manage TDC and BDC of the piston when starting the engine. The displacement is 26 mm smaller than the maximum stroke of 28 mm, due to the reserve for inertial forces and mechanical errors. From the geometric method, design parameters can be shown in Fig. 6.

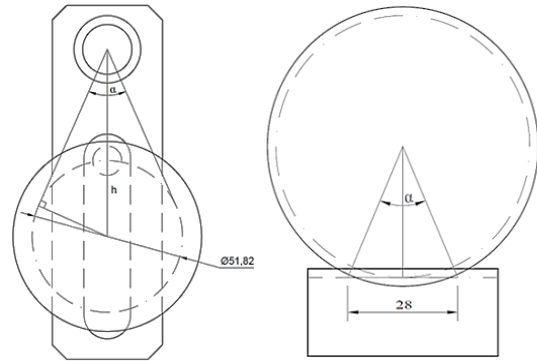


Fig. 6. The rack pinion mechanism of starting system.

#### V. RESULTS AND DISSCUSSION

The start-up process must ensure that sufficiently high pressure is maintained to allow the fuel to ignite. Thus, the engine must reach high speed and large torque to overcome friction, compression pressure at both ends of the cylinder. The experimental model of the mechanical starting system is established in Fig. 7.

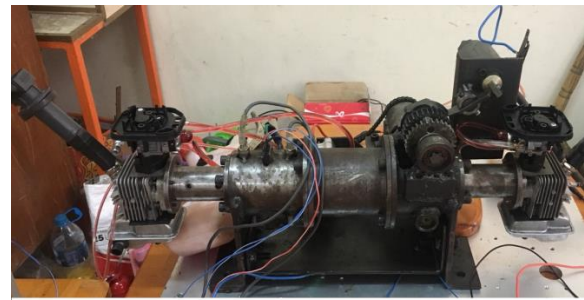


Fig. 7. Experimental model of PFE

Fig. 7 shows the starting system mechanism, the DC motor can achieve a maximum speed of 0.4 m/s with a frequency of about 5 Hz. However, the speed can be changed by changing gear size or gear ratio between the number of teeth on the driven and drive gear.

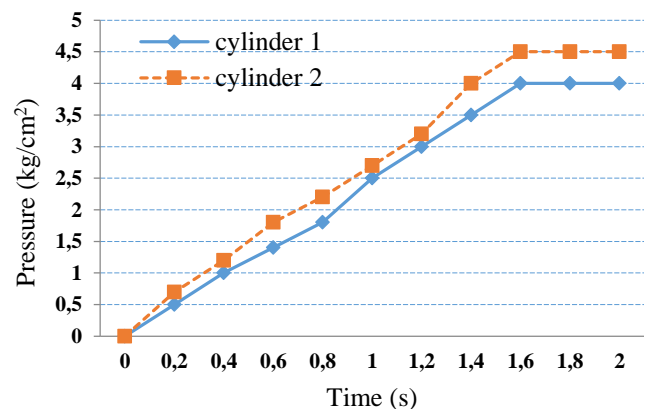


Fig. 8. Compression pressure in the cylinder and starting time

Fig. 9. illustrates the maximum compression pressure in the cylinder reaches 3.5 - 4.5 kg/cm<sup>2</sup> and stable for the rest of

the time, although the engine speed reaches its maximum in about 0.5 seconds, the maximum pressure is reached in about 2 seconds. That shows the maximum compression pressure in a cylinder must have a certain number of oscillations. The maximum pressure between cylinder 1 and cylinder 2 is not equal, which is predictable due to the inertial force in two cylinders which is difficult to control.

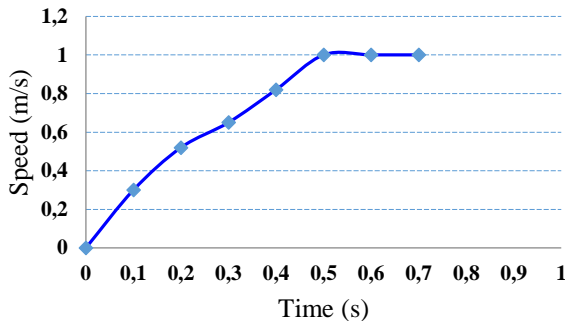


Fig. 9. Piston speed versus starting time

The engine prototype has successfully start up. However, the engine was not working stable, the reason for this is that the probability of misfiring at high engine speed. Misfiring may represent a problem in the free-piston engine since it does not have energy storage capable of driving the engine for several revolutions like the flywheel in a conventional engine [2]. Fig. 10 shows the maximum pressure when successful ignition reaches over 9 (kg/cm<sup>2</sup>) when the starting system is still working.

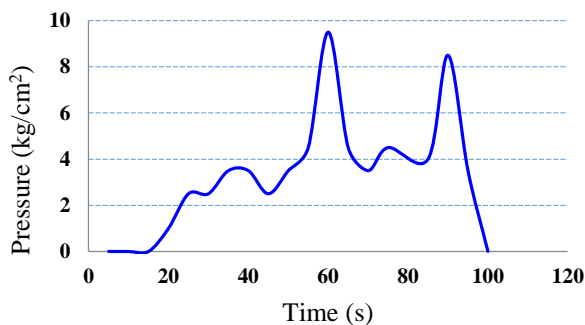


Fig. 10. Peak pressure for successful startup

## VI. CONCLUSIONS

The paper proposes a simple prototype of FPE with a mechanical starting system. Some preliminary results show that the successful engine startup can be able to obtain at a mean piston speed of 0.4 m/s and a maximum pressure of 4-4.5 (kg/cm<sup>2</sup>), the time taken to reach the peak pressure is relatively short in the range of 2 s. The DC motor drive the main shaft of FPE to startup at a stable speed. The pressure at the different cylinder heads is a problem that needs to be solved. The prototype has started a forced ignition test during start-up and has succeeded in a few cycles but requires more optimal research.

A potential challenge with the free-piston design is the engine control strategy. The compression stroke is driven by the power stroke expansion in the opposite cylinder. Variations in the combustion progress in one cylinder will influence the compression and combustion in the other. For the free-piston concept to be feasible, a control system which

can secure stable and smooth engine operation must be realized. In the future, lots of additional works must be performed before the FPE prototype comes true.

## ACKNOWLEDGMENT

This work belongs to the project grant No: B2019-SPK-08 funded by Ministry of Education and Training, and hosted by Ho Chi Minh City University of Technology and Education, Vietnam.

## REFERENCES

- [1] H. Feng, C. Guo, C. Yuan, Y. Guo, Z. Zuo, A. P. Roskilly, et al., "Research on combustion process of a free piston diesel linear generator," *Applied Energy*, vol. 161, pp. 395-403, 2016.
- [2] R. Mikalsen and A. P. Roskilly, "A review of free-piston engine history and applications," *Applied Thermal Engineering*, vol. 27, pp. 2339-2352, Oct 2007.
- [3] J. Hansson and M. Leksell, "Performance of a Series Hybrid Electric Vehicle with a Free-Piston Energy Converter," *IEEE Xplore*, 2007.
- [4] B. Jia, R. Mikalsen, A. Smallbone, and A. Paul Roskilly, "A study and comparison of frictional losses in free-piston engine and crankshaft engines," *Applied Thermal Engineering*, 2018.
- [5] M. R. Hanipah, R. Mikalsen, and A. P. Roskilly, "Recent commercial free-piston engine developments for automotive applications," *Applied Thermal Engineering*, vol. 75, pp. 493-503, 2015.
- [6] Nguyen Ba Hung, Ock Taeck Lim, "A study of a two-stroke free piston linear engine using numerical analysis," *Journal of Mechanical Science and Technology*, pp.1545-1557, 2014
- [7] Jia B, Tian G, Feng H, Zuo Z, Roskilly AP. "An experimental investigation into the starting process of free-piston engine generator," *Apply Energy* 2015;157: 798–804.
- [8] Q. Li, J. Xiao, and Z. Huang, "Simulation of a two-stroke free-piston engine for electrical power generation," *Energy & fuels*, vol. 22, pp. 3443-3449, 2008.
- [9] R. Mikalsen and A. P. Roskilly, "The control of a free-piston engine generator. Part 1: Fundamental analysis," *Applied Energy*, pp. 1273-1280, 2010
- [10] R. Mikalsen and A. P. Roskilly, "The control of a free-piston engine generator. Part 2: Engine dynamics and piston motion control," *Applied Energy*, pp. 1281-1287, 2010.
- [11] Boru Jia, Zhengxing Zuo, Huihua Feng, Guohong Tian, A. P. Roskilly, "Development approach of a spark-ignited free-piston engine generator," *SAE Technical Paper*, No. 2014-01-2894, 2014.
- [12] Blair, G. P. *Design and Simulation Two-stroke Engine*. Society of Automotive Engineers, Inc., 1996.
- [13] J. B. Heywood, *Internal combustion engine fundamentals*, McGraw-Hill Book Company, New York, USA (2018).
- [14] Fredriksson J, Bergman M, Golovitchev V, Denbratt I. "Modeling the effect of injection schedule change on free piston engine operation," *SAE paper* 2006-01- 0449; 2006.
- [15] Yitong Jiang a, Daojing Wang, Fushui Liu, Shuwei Zhao, and Qing Yang. "Dynamic Simulation of a Two-Stroke Spark Ignition Free-Piston Engine Generator,". pp 225-230. 2014.
- [16] R. Mikalsen, E. Jones, and A. P. Roskilly, "Predictive piston motion control in a free-piston internal combustion engine," *Applied Energy*, vol. 87, pp. 1722-1728, 2010.
- [17] R. Mikalsen, A.P. Roskilly, "The design and simulation of a twostroke free-piston compression ignition engine for electrical power generation," *Applied Thermal Engineering* (2007).
- [18] Jaeheun Kim, Choongsik Bae, Gangchul Kim, "Simulation on the effect of the combustion parameters on the piston dynamics and engine performance using the Wiebe function in a free piston engine," *Applied Energy*, pp. 446-455, 2014.
- [19] Johnson, T., Leick, M., and Moses, R., "Experimental Evaluation of a Prototype Free Piston Engine - Linear Alternator (FPLA) System," *SAE Technical Paper* 2016-01-0677, 2016.

# Numerical Study on Gas-Assisted Mold Temperature Control with the Application of Air Cover for Improving the Cavity Temperature in Injection Molding Process

Pham Son Minh  
HCMC University of Technology and  
Education, Ho Chi Minh City, 71307,  
Vietnam  
Ho Chi Minh City 72506, Vietnam  
minhps@hcmute.edu.vn

Tran Minh The Uyen  
HCMC University of Technology and  
Education, Ho Chi Minh City, 71307,  
Vietnam  
Ho Chi Minh City 72506, Vietnam  
uyentmt@hcmute.edu.vn

Thanh Trung Do  
HCMC University of Technology and  
Education, Ho Chi Minh City, 71307,  
Vietnam  
Ho Chi Minh City 72506, Vietnam  
trungtdt@hcmute.edu.vn

**Abstract**— In this paper, the gas-assisted mold temperature control was simulated with different cover characteristics for improving the cavity temperature in injection molding process. The heating was achieved with a stamp with the size of  $50 \times 50 \times 2$  mm. The model was meshed and simulated by ANSYS software. All boundary conditions were set up as the real experiment condition. The results showed that the sprue height (h) and the air height (V) affected the heating efficiency. The increase in sprue height from 3 mm to 7 mm would achieve the highest stamp temperature with the height of 3 mm. The same with sprue height, the change of air height (V) also has a position influence. In general, the heating with the cover assisted could increase the stamp temperature above  $300^\circ\text{C}$  within 20 s.

**Keywords**— Injection molding, dynamic mold temperature control, sprue height, air height, temperature distribution.

## I. INTRODUCTION

Presently, customer demands for higher product quality and more complex product geometries have necessitated the improvements of the existing techniques in injection molding. Mold temperature control is one of the effective methods used to address the issues encountered during the melt filling step [1–3]. For instance, molding of microproducts or parts with thin walls requires a high mold temperature. To meet the requirement of a high mold temperature during filling, substantial research has been conducted on many mold temperature control methods with the aim of increasing the cavity temperature while retaining a short cycle time. In these research, the most popular solution was preheating the cavity surface prior to melt filling. This step raises the surface temperature higher than the glass transition temperature ( $T_g$ ) of amorphous polymers or the melting point of semi crystalline polymers; thus, with high cavity temperature, the speed at which the frozen layer develops, as well as the melt viscosity and flow resistance, is reduced, allowing for an easier flow of the melt into the narrow cavity [4].

In the field of dynamic mold temperature control for injection molding process, the gas-assisted mold temperature control (GMTC) has primarily been applied for directly heating the mold surface [5, 6]. In this method, the hot gas from heating device is flown into the mold cavity. Heat

convection between the hot gas and cavity surface increases the energy received by the mold surface, thereby raising its temperature. The inlet and outlet of hot gas have to be inserted into the mold structure. Primary research conducted for GMTC revealed that the temperature of cavity surface could be increased to over  $150^\circ\text{C}$  and, with a proper design for the heating system, showed its ability to heat the complex cavity. However, one of the biggest challenges of using this method is that the mold structure must be redesigned relative to the requirements of the heating system, which often creates a highly complex mold structure for controlling the hot air flow. Additionally, recent research has shown that cavity temperature distribution with the gas-assisted mold temperature control is not very good that is the difference between temperatures could reach over  $60^\circ\text{C}$  for a heating area of  $80 \times 40$  mm. Therefore, the application of this method for the plastic injection mold process still needs further studies.

In this study, the gas-assisted mold temperature control in injection molding is employed for the mold temperature control within a gas temperature range from  $200^\circ\text{C}$  to  $400^\circ\text{C}$ . The general heating process consists of three steps, as follows:

- Step 1: The hot gas source is transported to the heating area at which the cavity temperature is still low.
- Step 2: The hot gas ejects directly into the heating area and the cavity surface temperature will rise.
- Step 3: The gas drier is removed from the molding area and two half molds are closed before starting a new molding cycle.

Moreover, a numerical study will be achieved for observing the heating efficiency of this type of process for injection molding cycle.

## II. SIMULATION METHOD

In this study, the GMTC heating process was analyzed by using the ANSYS software. The numerical model as an air cover was designed to analyze the distribution of the mold surface temperature in injection molding process. The material of air cover was aluminum with density of  $2700\text{ kg/m}^3$ . Based on the structure of cavity of mold, the



insulation component was used to cover the heating area so that simulation models included only two parts: the insert and hot gas part. The model and boundary conditions of numerical simulation are presented in Fig.1. For increasing the simulation precision, the mesh of part insert was generated with the hex-dominant elements. The inflation meshing method was applied with 10 layers at the contact surfaces. Moreover, the tetrahedron elements were used for the air volume. The temperatures of gas inlet used for the simulation ranged from 200 °C to 400 °C. The simulation parameters were set up as in the experiments and showed in Table I.

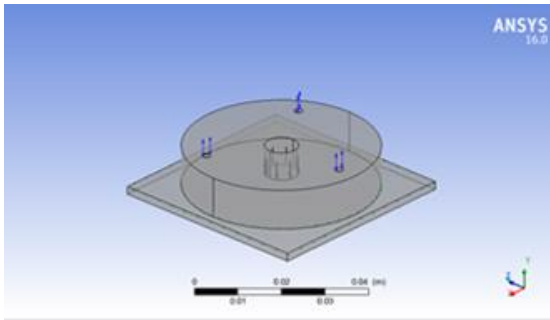


Fig. 1. Simulation model

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Hot air temperature	200 – 400 °C
Density of hot air [33]	1.17 kg/m <sup>3</sup>
Heat capacity of hot air [33]	1006 J/kg·K
Thermal-expansion coefficient of hot air [33]	3.4 × 10 <sup>-3</sup> /K
Air pressure	0.1 MPa
Aluminum density – ASTM B209-14	2700 kg/m <sup>3</sup>

### III. RESULTS AND DISCUSSION

The dynamic mold temperature control was studied to increase the cavity temperature of mold in injection molding process that was demonstrated in previous studies [7, 8, 9]. However, it was still not suitable in some cases due to the different temperatures at the heating surface for different models. In this research, the temperature distribution on insert plate was studied to estimate the heating efficiency of GMTC with the application of air cover in the mold heating process. The gas drier had one hot gas inlet that mainly served to heat the insert plate and improve the temperature distribution as well as the heating rate. The heating step was analyzed with the meshing model in Fig. 2 with the hot air temperature of 400 °C and heating time of 20 s; various types of air covers were used to find out the best shape for heating process.

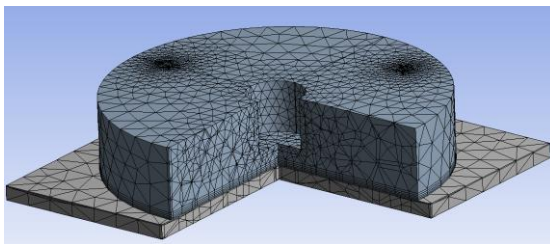
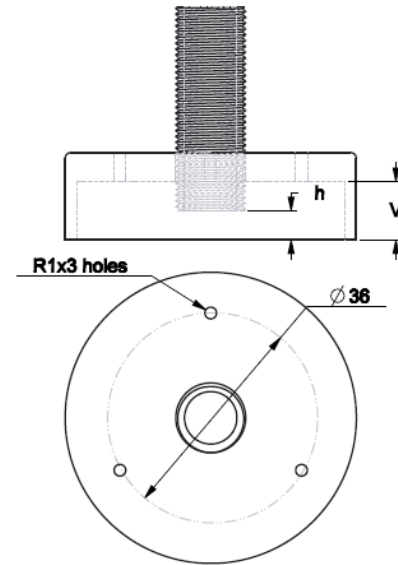
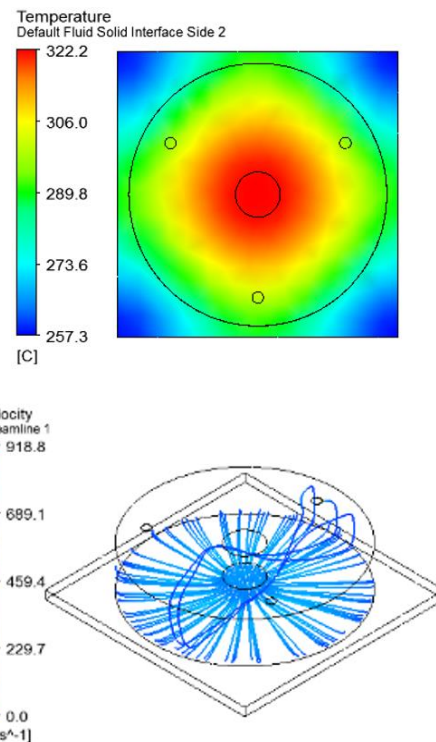


Fig. 2. Meshing model

Fig. 3 shows the experiment model for estimating the GMTC with the cover. In this paper, the cover was designed for heating a circle area with the diameter of 42 mm. The

inlet of hot air was sprayed at the top, and the outlet was designed with 3 holes at the top of the cover. The hot air was flowed through a pipe with the diameter of 8 mm. The hot air outlets have the diameter of 2.0 mm. In our previous research, the gap between the heating surface and the heating source has a strong influence on the heating efficiency as well as the temperature distribution [9, 10]. Therefore, in this research, the distance from the end of this pipe to the heating surface is  $h$ , which will be varied from 3 mm to 7 mm. In addition, in this paper, for focusing the thermal energy on the heating surface, the heating cover was designed and applied for the heating process, which will create the air volume with high temperature. For observing the effect of this hot air volume on the heating efficiency, the volume of hot air ( $V$ ) will be changed from 5 mm to 15 mm.

Fig. 3. Air height ( $V$ ) and sprue height ( $h$ )Fig. 4. Temperature distribution with the Sprue height ( $h$ ) of 3 mm



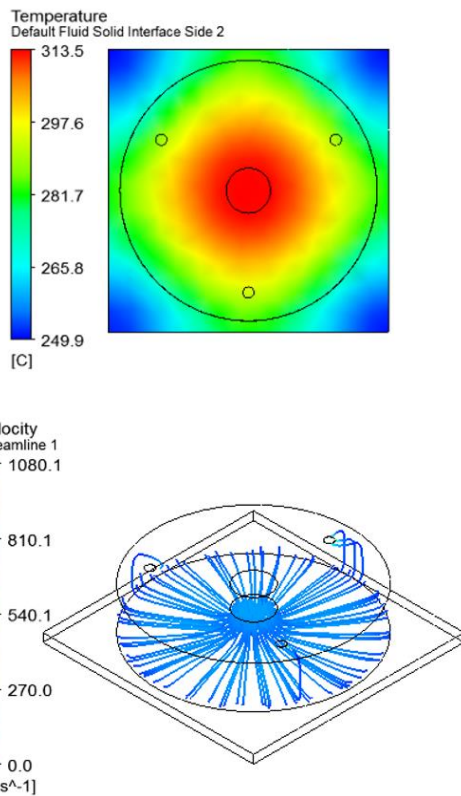


Fig. 5. Temperature distribution with the Sprue height (h) of 5 mm

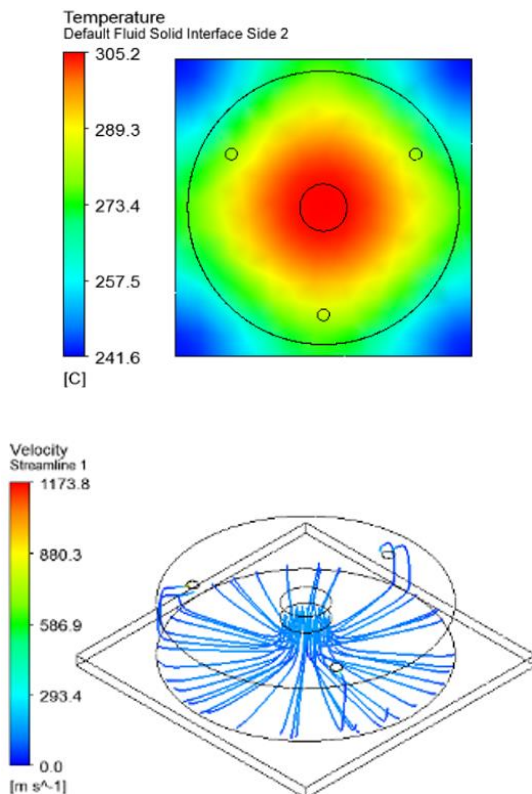


Fig. 6. Temperature distribution with the Sprue height (h) of 7 mm

By simulation, Figures 4, 5 and 6 present the temperature distribution of the insert surface by different injection distance, with height of air volume  $V = 10$  mm and three outlet holes on top of the air cover. This result shows that the highest temperature was reached with the injection distance of 3 mm (322 °C). Meanwhile, the lowest temperature is 305

°C with the injection distance of 7 mm. Furthermore, with the injection distance of 7 mm, the air streamline got the largest value at  $1173.8 \text{ ms}^{-1}$  and had the best temperature distribution on the insert plate. However, with the injection distance of 3 mm, the air streamline got the smallest value at  $918.8 \text{ ms}^{-1}$  and the worst temperature distribution. From the simulation results, it can be seen that the injection distance greatly affects the temperature and thermal distribution of the insert; the smaller the injection distance is, the higher the cavity temperature becomes, because small injection distance brings the greater heat absorption capacity of the insert plate. Moreover, a longer injection distance can lead to the better temperature distribution due to the hot air moving throughout the cover before escaping, so the velocity of air flow is high.

The simulation results given in Figures 7, 8 and 9 also illustrate that the height of air volume has effects on the heating rate and temperature distribution. With the height of air volume ranging from  $V = 5$  mm to 15 mm and injection distance  $h = 3$  mm, it can be seen that with  $V = 5$  mm, the temperature reached the maximum value of 332.2 °C but had the worst temperature distribution. However, with  $V = 15$  mm, although it got the smallest temperature (310.5 °C) and air streamline value ( $807.3 \text{ m s}^{-1}$ ), this kind of air cover reached the best temperature distribution. In addition, the largest air streamline value achieved with  $V = 10$  mm. Therefore, the results show that the height of air volume had a significant effect on the temperature and thermal distribution of the insert. The smaller the height of air volume is, the higher the cavity temperature becomes, because a smaller height of air volume can lead to the better heat absorption capacity of the insert plate, whereas a greater height of air volume can get the better temperature distribution. In addition, because of the longer hot air condensation in the cover before escaping, the hot air will transmit heat to lower temperature regions.

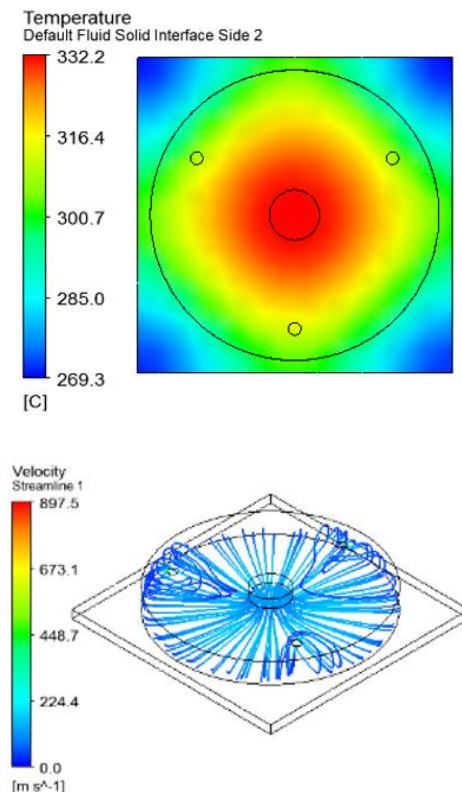


Fig. 7. Temperature distribution with the air height (V) of 5 mm.

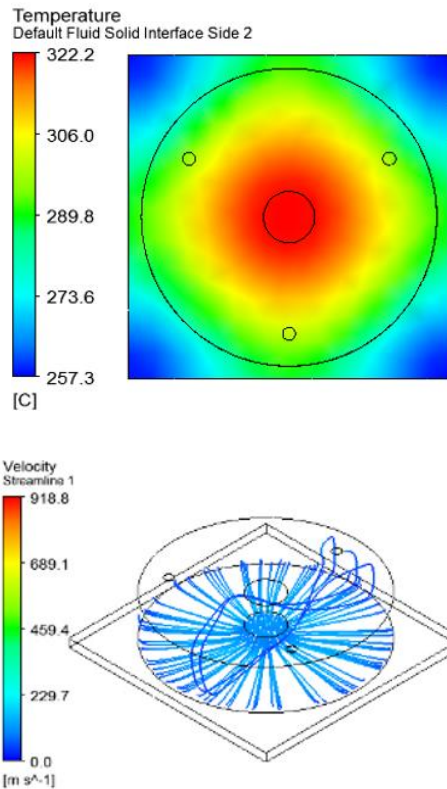


Fig. 8. Temperature distribution with the air height (V) of 10 mm.

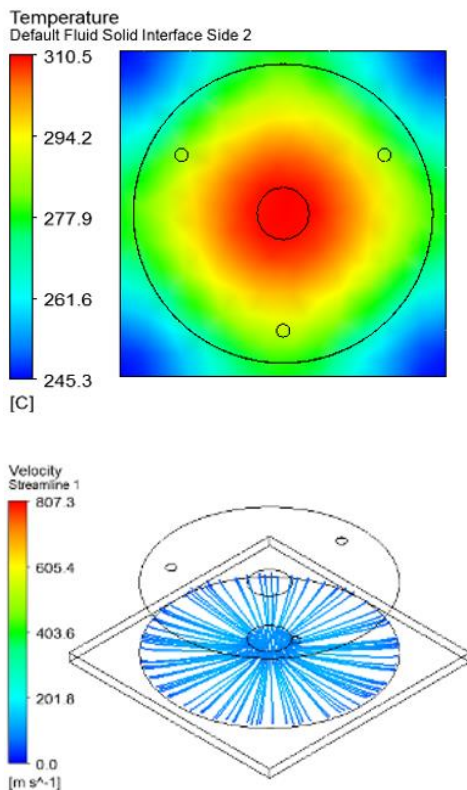


Fig. 9. Temperature distribution with the air height (V) of 15 mm.

#### IV. CONCLUSION

In this study, separate simulations and experiments were conducted to evaluate the temperature distribution by using air cover in heating process and its application in improving melt flow length of injection molding products. The significant findings could be generalized as follows:

- In the of sprue height (h) ranging from 3 mm to 7 mm, the case of 3 mm results in the highest temperature, and 7 mm results in the lowest temperature.
- With the case of air heights (V), the best temperature distribution on the surface of insert appeared with the air height of 5 mm.
- With the heating time of 20 s, the GMTC could increase the stamp temperature to over 300 °C. This result satisfies almost the requirement of dynamic mold temperature control in injection molding cycle.

#### ACKNOWLEDGMENT

This work belongs to the project grant No: B2020-SPK-01 funded by Ministry of Education and Training, and hosted by Ho Chi Minh City University of Technology and Education, Vietnam.

#### REFERENCES

- [1] P. Zhao, J. Zhang, Z. Dong, J. Huang, H. Zhou, J. Fu, and L-S. Turng, "Intelligent Injection Molding on Sensing, Optimization, and Control," *Adv. Polym. Technol.*, vol. 2020, pp. 1–22, 2020.
- [2] R. Spina, "Technological characterization of PE/EVA blends for foam injection molding," *Mater. Des.*, vol. 84, pp. 64–71, 2015.
- [3] M-C. Jeng, S-C Chen, P. S. Minh, J-A Chang, and C. Chung, "Rapid mold temperature control in injection molding by using steam heating," *Int. Commun. Heat. Mass. Transf.*, vol. 37, pp. 1295–1304, 2010.
- [4] M. Hammami, F. Kria, and M. Baccar, "Numerical study of thermal control system for rapid heat cycle injection molding process," *P. I. Mech. Eng. E-J. Pro.*, pp. 1–15, 2015.
- [5] S-C Chen, P. S. Minh, and J. A. Chang, "Gas-assisted mold temperature control for improving the quality of injection molded parts with fiber additives," *Int. Commu. Heat. Mass.*, vol. 38, pp. 304–312, 2011.
- [6] S. C. Chen, R. D. Chien, S. H. Lin, M. C. Lin, and J. A. Chang, "Feasibility evaluation of gas-assisted heating for mold surface temperature control during injection molding process," *Int. Commu. Heat. Mass.*, vol. 36, 806–812, 2009.
- [7] S. Menotti, H. N. Hansen, G. Bissacco, M. Calao, P. T Tang, and C. Ravn, "Injection molding of nanopatterned surfaces in the sub-micrometer range with induction heating aid," *Int. J. Adv. Manuf. Tech.*, vol. 2014, pp.1–10, 2014.
- [8] P. Guerrier, G. Tosello, K. Nielsen, J. H. Hattel, "Three-dimensional numerical modeling of an induction heated injection molding tool with flow visualization," *Int. J. Adv. Manuf. Technol.*, vol. 2015, pp. 1–18, 2015.
- [9] Phan The Nhan, Thanh Trung Do, Tran Anh Son and Pham Son Minh, "Study on external gas-assisted mold temperature control for improving the melt flow length of thin rib products in the injection molding process", *Advances in Polymer Technology*, Vol. 2019, Article ID 5973403, pp. 1-17.
- [10] Tran Minh The Uyen, Nguyen Truong Giang, Thanh Trung Do, Tran Anh Son and Pham Son Minh, "External gas-assisted mold temperature control improves weld line quality in the injection molding process," *Materials* 2020, 13, 2855; doi:10.3390/ma13122855.

# PROCEEDINGS OF

## 2020 5<sup>TH</sup> INTERNATIONAL CONFERENCE ON GREEN TECHNOLOGY AND SUSTAINABLE DEVELOPMENT (GTSD)

This work belongs to the contract number 105/2020-HĐ-  
SKHCN funded by Ho Chi Minh City Department of Science  
and Technology, Vietnam

