

Lecture Notes in Networks and Systems 570

Leonard Barolli *Editor*

Advances on Broad-Band Wireless Computing, Communication and Applications

Proceedings of the 17th International
Conference on Broad-Band Wireless
Computing, Communication and
Applications (BWCCA-2022)

 Springer

Lecture Notes in Networks and Systems

Volume 570

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA,
School of Electrical and Computer Engineering—FEEC, University of Campinas—
UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering,
Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University
of Illinois at Chicago, Chicago, USA

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of
Alberta, Alberta, Canada

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering,
KIOS Research Center for Intelligent Systems and Networks, University of Cyprus,
Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong,
Kowloon, Hong Kong

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

More information about this series at <https://link.springer.com/bookseries/15179>

Leonard Barolli
Editor

Advances on Broad-Band Wireless Computing, Communication and Applications

Proceedings of the 17th International
Conference on Broad-Band Wireless
Computing, Communication and Applications
(BWCCA-2022)

 Springer

Editor

Leonard Barolli
Department of Information and
Communication Engineering
Fukuoka Institute of Technology
Fukuoka, Japan

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-031-20028-1

ISBN 978-3-031-20029-8 (eBook)

<https://doi.org/10.1007/978-3-031-20029-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Welcome Message of BWCCA-2022 International Conference Organizers

Welcome to the 17th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA-2022), which will be held in conjunction with 3PGCIC-2022 International Conference from October 27 to October 29, 2022.

This International Conference is a forum for sharing ideas and research work in the emerging areas of broadband and wireless computing. Information networks of today are going through a rapid evolution. Different kinds of networks with different characteristics are emerging, and they are integrating in heterogeneous networks. For these reasons, there are many interconnection problems which may occur at different levels of the hardware and software design of communicating entities and communication networks. These kinds of networks need to manage an increasing usage demand, provide support for a significant number of services, guarantee their QoS, and optimize the network resources.

The success of all-IP networking and wireless technology has changed the ways of the people living around the world. The progress of electronic integration and wireless communications is going to pave the way to offer people the access to the wireless networks on the fly, based on which all electronic devices will be able to exchange the information with each other in ubiquitous way whenever necessary.

The aim of this conference is to present the innovative research and technologies as well as developments related to broadband networking and mobile and wireless communications.

The organization of an International Conference requires the support and help of many people. A lot of people have helped and worked hard to produce a successful BWCCA-2022 technical program and conference proceedings. First, we would like to thank all Authors for submitting their papers, Program Committee Members, and Reviewers who carried out the most difficult work by carefully evaluating the submitted papers.

We thank Web Administrators Co-Chairs and Finance Chair for their excellent work. We would like to express our gratitude to Prof. Makoto Takizawa, Hosei University, Japan, as Honorary Chair of BWCCA-2022 for his support and

help. We give special thanks to Keynote Speakers of BWCCA-2022 and local arrangement team.

We hope you will enjoy the conference proceedings.

BWCCA-2022 Organizing Committee

Honorary Chair

Makoto Takizawa

Hosei University, Japan

General Co-chairs

Vladi Kolic

Polytechnic University of Tirana, Albania

Tomoya Enokido

Rissho University, Japan

Hsing-Chung Chen

Asia University, Taiwan

Program Committee Co-chairs

Evjola Spaho

Polytechnic University of Tirana, Albania

Naohiro Hayashibara

Kyoto Sangyo University, Japan

Hyunhee Park

Myongji University, Korea

International Advisory Committee

Fang-Yie Leu

Tunghai University, Taiwan

David Taniar

Monash University, Australia

Kangbin Yim

SCH University, Korea

Publicity Co-chairs

Lidia Ogiela

AGH University of Science and Technology,
Poland

Keita Matsuo

Fukuoka Institute of Technology, Japan

Tetsuya Shigeyasu

Prefectural University of Hiroshima, Japan

Finance Chair

Makoto Ikeda Fukuoka Institute of Technology, Japan

Web Administrator Chairs

Phudit Ampririt Fukuoka Institute of Technology, Japan
 Kevin Bylykbashi Fukuoka Institute of Technology, Japan
 Ermioni Qafzezi Fukuoka Institute of Technology, Japan

Local Organizing Co-chairs

Aleksander Biberaj Polytechnic University of Tirana, Albania
 Ilir Shinko Polytechnic University of Tirana, Albania
 Bexhet Kamo Polytechnic University of Tirana, Albania

Steering Committee Chair

Leonard Barolli Fukuoka Institute of Technology, Japan

Track Areas

Next Generation Wireless Networks

Track Co-chairs

Bhed Bista Iwate Prefectural University, Japan
 Szu-Yin Lin Chung Yuan Christian University, Taiwan
 Sriram Chellappan University of South Florida, USA

PC Members

Jiahong Wang Iwate Prefectural University, Japan
 Shigetomo Kimura University of Tsukuba, Japan
 Chotipat Pornavalai King Mongkut's Institute of Technology
 Ladkrabang, Thailand
 Danda B. Rawat Howard University, USA
 Gongjun Yan University of Southern Indiana, USA
 Vamsi Paruchuri University of Central Arkansas, USA
 Arjan Durrezi IUPUI, USA
 Shih-Yi James Chien National Sun Yat-sen University, Taiwan
 Pei-Ju Lee National Chung Cheng University, Taiwan
 Chih-Hao Lin Chung Yuan Christian University, Taiwan
 Hao-Hsiang Ku National Taiwan Ocean University, Taiwan
 Jung-Bin Li Fu Jen Catholic University, Taiwan

Thoshitha Gamage	Southern Illinois University, USA
Mukundan Sridharan	Samraksh Company, USA
Brijesh Chejerla	Florida Blue, USA
Srinivas Chakravarthi Thandu	Amazon, USA

Cloud and Service Computing

Track Co-chairs

Hwamin Lee	Soonchunhyang University, Korea
Ramesh C. Hansdah	Indian Institute of Science, Bangalore, India
Baojiang Cui	Beijing University of Posts and Telecommunications, China

PC Members

Gang Wang	Nankai University, China
Jianxin Wang	Beijing Forestry University, China
Jie Cheng	Shandong University, China
Shaoyin Cheng	University of Science And Technology of China, China
Yan Zhang	Hubei University, China
Willy Susilo	University of Wollongong, Australia
Kamil Kluczniak	Wroclaw University of Technology, Poland
Francesco Palmieri	University of Salerno, Italy
Jian Shen	Nanjing University of Information Science and Technology, China
Jin Li	Guangzhou University, China
Fanguo Zhang	Sun Yat-sen University, China
Xinyi Huang	Fujian Normal University, China
Shengli Liu	Shanghai Jiaotong University, China
Zhenjie Huang	Zhangzhou City University, China
Joseph K. Liu	Institute for Infocomm Research, Australia
Yong Yu	University of Wollongong, China
Ding Wang	Peking University, China
Tao Jiang	Xidian University, China
Jianfeng Wang	Xidian University, China
S. D. Madhu Kumar	NIT, Calicut, India
Ashutosh Bhatia	BITS Pilani, Pilani Campus, India
Amulya Rathna Swain	KIIT, Bhubaneshwar, India
Yogesh Simmhan	IISc, Bangalore, India
Soumya K. Ghosh	Indian Institute of Technology, India

Multimedia and Web Applications

Track Co-chairs

Yoshihiro Okada	Kyushu University, Japan
Chuan-Yu Chang	National Yunlin University of Science and Technology, Taiwan
Salem Alkhalaf	Qassim University, Saudi Arabia

PC Members

Kaoru Sugita	Fukuoka Institute of Technology, Japan
Tomoyuki Ishida	Fukuoka Institute of Technology, Japan
Makoto Nakashima	Oita University, Japan
Nobukazu Iguchi	Kinki University, Japan
Kenzi Watanabe	Hiroshima University, Japan
Shinji Sugawara	Chiba Institute of Technology, Japan
Li-Wei Kang	National Yunlin University of Science and Technology, Taiwan
Chia-Hung Yeh	National Taiwan Normal University, Taiwan
Jun-Wei Hsieh	National Taiwan Ocean University, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Chien-Cheng Lee	Yuan-Ze University, Taiwan
Muhammad Hussain	King Saud University, Saudi Arabia
Umair Azfar Khan	Habib University, Pakistan
Shigeru Takano	Kyushu University, Japan
Kosuke Kaneko	Kyushu University, Japan
Akira Haga	Kyushu University, Japan
Wei Shi	Kyushu University, Japan

Security and Privacy

Track Co-chairs

Tianhan Gao	Northeastern University, China
Masakatsu Nishigaki	Shizuoka University, Japan
Mohamed Abdur Rahman	Prince Mughrin University, Saudi Arabia

PC Members

Nan Guo	Northeastern University, China
Zhenhua Tan	Northeastern University, China
Jian Xu	Northeastern University, China
Hiroaki Kikuchi	Meiji University, Japan

Takamichi Saito	Meiji University, Japan
Rashid Tahir	University of Prince Mugrin Madinah, Saudi Arabia
Syed Sadiq	University of Prince Mugrin Madinah, Saudi Arabia
Md. Mamunur Rashid (Mamun)	King's Business School, UK
Akhlaq Ahmad	Umm Al Qura University Makkah, Saudi Arabia
Shyhtsun Felix Wu	University of California, Davis, USA
Zhen-Yu Wu	Penghu University of Science and Technology, Taiwan
Tsung-Chih Hsiao	Southeast University, China
Kuo-Kun Tseng	Harbin Institute of Technology, China
Akira Otsuka	Institute of Information Security, Japan
Naonobu Okazaki	University of Miyazaki, Japan
Masaki Shimaoka	Secom Co., Ltd., Japan
Davinder Kaur	IUPUI, USA

Network Protocols and Performance Analysis

Track Co-chairs

Tetsuya Shigeyasu	Prefectural University of Hiroshima, Japan
Ching-Feng Liang	Industrial Technology Research Institute, Taiwan
Vamsi Paruchuri	University of Central Arkansas, USA

PC Members

Xiaoyi Wang	Nokia Solutions and Networks, USA
Yu Sun	University of Central Arkansas, USA
Qiang Duan	Pennsylvania State University, USA
Han-Chieh Wei	Dallas Baptist University, USA
Masaaki Yamanaka	Japan Coast Guard Academy, Japan
Misako Urakami	Tokuyama College of Technology, Japan
Tomoya Kawakami	Nara Institute of Science and Technology, Japan
Masaaki Noro	Fujitsu Corp., Japan
Nobuyoshi Sato	Iwate Prefectural University, Japan
Phone Lin	National Taiwan University, Taiwan
Ray-Guang Cheng	National Taiwan University of Science and Technology, Taiwan
Shun-Ren Yang	National Tsing Hua University, Taiwan
Whai-En Chen	National Ilan University, Taiwan

Intelligent and Cognitive Computing

Track Co-chairs

Lidia Ogiela	AGH University of Science and Technology, Poland
Takahiro Uchiya	Nagoya Institute of Technology, Japan
Hai Dong	RMIT University, Australia

PC Members

Atsuko Mutoh	Nagoya Institute of Technology, Japan
Shinsuke Kajioka	Nagoya Institute of Technology, Japan
Ryota Nishimura	Tokushima University, Japan
Shohei Kato	Nagoya Institute of Technology, Japan
Francesco Pascale	University of Salerno, Italy
Jan Platoš	VŠB Technical University of Ostrava, Czech Republic
Pavel Krömer	VŠB Technical University of Ostrava, Czech Republic
Urszula Ogiela	AGH University of Science and Technology, Poland
Jana Nowaková	VŠB Technical University of Ostrava, Czech Republic
Chang Choi	Chosun University, Korea
Hoon Ko	Chosun University, Korea
Hae-Duck Joshua Jeong	Korean Bible University, Korea
Pengcheng Zhang	Hohai Univesity, China
Sajib Mistry	University of Sydney, Australia
Tooba Aamir	RMIT University, Australia
Wei Du	Wuhan University of Technology, China
Wei Zhang	Macquarie University, Australia
Shang-Pin Ma	National Taiwan Ocean University, Taiwan

Distributed and Parallel Computing

Track Co-chairs

Naohiro Hayashibara	Kyoto Sangyo University, Japan
Omar Khadeer Hussain	University of New South Wales (UNSW), Australia

PC Members

Sazia Parvin	Melbourne Polytechnic, Australia
Naeem Janjua	Edith Cowan University, Australia
Alireza Faed	Toronto Metropolitan University, Canada

Adil Hammadi	Curtin University, Australia
Lucian Prodan	Polytechnic University Timisoara, Romania
Kanwalinderjit Kaur Gagneja	Florida Polytechnic University
Rohaya Latip	Universiti Putra Malaysia, Malaysia
Tomoya Enokido	Rissho University, Japan
Makoto Takizawa	Hosei University, Japan
Leonard Barolli	Fukuoka Institute of Technology, Japan
Akio Koyama	Yamagata University, Japan
Minoru Uehara	Toyo University, Japan

IoT and Smart Environment

Track Co-chairs

Nadeem Javaid	COMSATS University Islamabad, Pakistan
Chun-Wei Tsai	National Chung Hsing University, Taiwan

PC Members

Zahoor Ali Khan	Higher Colleges of Technology, UAE
Umar Qasim	University of Engineering and Technology, Lahore, Pakistan
Farookh Hussain	University of Technology Sydney, Australia
Elis Kulla	Fukuoka Institute of Technology, Japan
Keita Matsuo	Fukuoka Institute of Technology, Japan
Hsin-Hung Cho	National Ilan University, Taiwan
Fan-Hsun Tseng	National Taiwan Normal University, Taiwan
Hsin-Te Wu	National Penghu University of Science and Technology, Taiwan

Database, Data Mining, and Big Data

Track Co-chairs

Antonio Esposito	University of Campania “Luigi Vanvitelli”, Italy
Yao-Chung Fan	National Chung Hsing University, Taiwan

PC Members

Mehran Samavati	University of Sydney, Australia
Farshid Hajati	Griffith University, Australia
Jinnie Hee Yoon	Sejong University, Korea
Elena Sitnikova	UNSW, Australia
Chen-Yi Lin	National Taichung University of Science and Technology, Taiwan

Lun-Chi Chen	National Center for High-performance Computing (NCHC), Taiwan
Huan Chen	National Chung Hsing University, Taiwan
Luca Tasquier	University of Campania “Luigi Vanvitelli”, Italy
Stefania Nacchia	University of Campania “Luigi Vanvitelli”, Italy
Salvatore Augusto Maisto	University of Campania “Luigi Vanvitelli”, Italy
Salvatore D’Angelo	University of Campania “Luigi Vanvitelli”, Italy

Ubiquitous and Pervasive Computing

Track Co-chairs

Isaac Woungang	Toronto Metropolitan University, Canada
Asm Kayes	La Trobe University, Australia
Chyi-Ren Dow	Feng Chia University, Taiwan

PC Members

Evjola Spaho	Polytechnic University of Tirana, Albania
Makoto Ikeda	Fukuoka Institute of Technology, Japan
Elis Kulla	Fukuoka Institute of Technology, Japan
Admir Barolli	Aleksander Moisiu University of Durres, Albania
Donald Elmazi	Canadian Institute of Technology, Albania
Alan Colman	Swinburne University of Technology, Australia
Iqbal H. Sarker	Swinburne University of Technology, Australia
Eric Pardede	La Trobe University, Australia
Syed Mahbub	La Trobe University, Australia
Patrick Hung	The University of Ontario Institute of Technology, Canada
Pei-Chun Lin	Feng Chia University, Taiwan
Zhang Kejun	ZheJiang University, China
Duc-Binh Nguyen	Thai Nguyen University of Information and Communications Technology (ICTU), Vietnam
Wei Lu	Keene State College, USA
Luca Caviglione	CNIT, Italy
Hamed Aly	Acadia University, Canada
Danda B. Rawat	Howard University, USA
Marcelo Luis Brocardo	University of Victoria, Canada
Glauco Carvalho	Ryerson University, Canada

BWCCA-2022 Reviewers

Barolli Admir
Barolli Leonard
Bista Bhed
Chellappan Sriram
Chen Hsing-Chung
Cui Baojiang
Di Martino Beniamino
Durresti Arjan
Enokido Tomoya
Fun Li Kin
Funabiki Nobuo
Gao Tianhan
Gotoh Yusuke
Hussain Farookh
Hussain Omar
Javaid Nadeem
Ikeda Makoto
Ishida Tomoyuki
Kanzaki Akimitsu
Kayem Anne
Kayes Asm
Kikuchi Hiroaki
Koyama Akio
Kulla Elis
Leu Fang-Yie

Matsuo Keita
Nishigaki Masakatsu
Ogiela Lidia
Ogiela Marek
Okada Yoshihiro
Paruchuri Vamsi Krishna
Rahayu Wenny
Sakamoto Shinji
Shibata Yoshitaka
Shigeyasu Tetsuya
Saito Takamichi
Spaho Eviola
Sugawara Shinji
Takizawa Makoto
Taniar David
Uehara Minoru
Venticinque Salvatore
Vitabile Salvatore
Waluyo Agustinus Borgy
Wang Xu An
Woungang Isaac
Xhafa Fatos
Yi Liu
Yim Kangbin

BWCCA-2022 Keynote Talks

Humanics Information Security: How to Go Above and Beyond?

Masakatsu Nishigaki

Shizuoka University, Hamamatsu City, Japan

Abstract. Who uses the information systems? The answer is, of course, human beings. Who attacks the information systems? The answer is, unfortunately, human beings again. Therefore, any system security that does not consider user characteristics (from viewpoints of both legitimate and malicious users) is pointless. The key is how to combine security technologies and human factors, specifically cognitive and psychological characteristics, in designing information systems. We call the concept “humanics information security”. In a digital transformation environment, information systems around us will be automatized, and artificial intelligence can automatically support our lives. However, we must not forget that an important decision should not be fully automated, but our consent is necessary when a critical decision is made. This means that human still remains even in extremely advanced automated information systems as its weakest link in information security. In this talk, we will discuss how the humanics information security approach can enhance both security and usability of information systems.

Multi-objective Methods for Wireless Sensor Network Optimization

Pavel Krömer

VSB-Technical University of Ostrava, Ostrava, Czech Republic

Abstract. When designing a wireless sensor network, several performance metrics should be considered, e.g., network lifetime, target coverage, and sensor energy consumption. Very often, these metrics are in conflict with each other, which means that by optimizing some of them, we worsen the others. Designing the network is therefore a problem of multi-objective optimization. In this talk, we provide an overview of selected multi-objective wireless sensor network design problems and outline several methods proposed to tackle them. Special attention is paid to the optimization of network lifetime and target coverage. We consider two variants of the algorithm, in which the fitness function comprises only the network lifetime, or where it includes both, the network lifetime and target coverage. This makes it possible to find a trade-off between these two objectives. The ability of multi-objective metaheuristics to tackle such problems is demonstrated on a genetic algorithm designed to solve this challenge.

Contents

A Comparison Study of UNDX and UNDX-m Methods for LDVM and RDVM Router Replacement Methods by WMN-PSODGA Hybrid Intelligent System Considering Stadium Distribution	1
Admir Barolli, Kevin Bylykbashi, Ermioni Qafzezi, Shinji Sakamoto, Leonard Barolli, and Makoto Takizawa	
Performance Comparison of Roulette Wheel and Random Selection Methods by WMN-PSODGA Simulation System Considering Stadium Distribution and LDIWM	15
Kevin Bylykbashi, Ermioni Qafzezi, Phudit Ampririt, Admir Barolli, Elis Kulla, and Leonard Barolli	
A Fuzzy-Based System for Handover in 5G Wireless Networks Considering Different Network Slicing Constraints: Effects of Slice Reliability Parameter on Handover Decision	27
Phudit Ampririt, Ermioni Qafzezi, Kevin Bylykbashi, Makoto Ikeda, Keita Matsuo, and Leonard Barolli	
A Simulated Annealing Based Simulation System for Optimization of Wild Deer Damage Prevention Devices	38
Sora Asada, Kyohei Toyoshima, Aoto Hirata, Yuki Nagai, Nobuki Saito, Tetsuya Oda, and Leonard Barolli	
Techno-Economic Analysis of Cloud Computing Supported by 5G: A Cloud vs on Premise Based Solutions Comparison	45
Christos Bouras, Charalampos Chatzigeorgiou, Anastasia Kollia, and Philippos Pouyioutas	
An Integrated Fog-VDTN Architecture for Data Dissemination	59
Evjola Spaho	

Energy-Consumption Evaluation of the Tree-Based Fog Computing (TBFC) Model	66
Dilawaer Duolikun, Shigenari Nakamura, Tomoya Enokido, and Makoto Takizawa	
Evaluation of the Information Flow Control in the Fog Computing Model	78
Shigenari Nakamura, Tomoya Enokido, and Makoto Takizawa	
A Study of Network Attack Strategy Using AS Topology Map	91
Naoya Sekiguchi and Hidema Tanaka	
Improving Classification Accuracy by Optimizing Activation Function for Convolutional Neural Network on Homomorphic Encryption	102
Kohei Yagyu, Ren Takeuchi, Masakatsu Nishigaki, and Tetsushi Ohki	
An Attention Mechanism for Visualizing Word Weights in Source Code of PowerShell Samples: Experimental Results and Analysis	114
Yuki Mezawa and Mamoru Mimura	
Improving Palmprint-Region Estimation for ID-Less Palmprint Recognition	125
Ayumi Serizawa, Ryosuke Okudera, Yumo Ouchi, Mizuho Yoshihira, Yuya Shiomi, Naoya Nitta, Masataka Nakahara, Akira Baba, Yutaka Miyake, Tetsushi Ohki, and Masakatsu Nishigaki	
Real Vehicle-Based Attack Dataset for Security Threat Analysis in a Vehicle	137
Yeji Koh, Yoonji Kim, Munkhdelgerekh Batzorig, and Kangbin Yim	
Performance Analysis of HARQ in 2-step RACH Procedure Using Markov Chain Model	147
Byungchan Kim and Hyunhee Park	
A Comparison Study of FC-RDVM with LDVM Router Replacement Methods by WMN-PSOHC Simulation System Considering Weibull Distribution of Mesh Clients	159
Shinji Sakamoto, Admir Barolli, Yi Liu, Elis Kulla, Leonard Barolli, and Makoto Takizawa	
A Fuzzy-Based System for Estimation of Landslide Disasters Risk Considering Digital Elevation Model	167
Kei Tabuchi, Kyohei Toyoshima, Nobuki Saito, Aoto Hirata, Yuki Nagai, Tetsuya Oda, and Leonard Barolli	
Human-Centered Protocols for Secure Data Management in Distributed Systems	175
Urszula Ogiela, Makoto Takizawa, and Lidia Ogiela	

Multi-Version Concurrency Control to Reduce the Electric Energy Consumption of Servers 180
 Tomoya Enokido, Dilawaer Duolikun, and Makoto Takizawa

A Study on Increasing Simultaneous Transmissions After Extended RTS/CTS Handshake on Full-duplex Wireless LANs 192
 Hikari Hashimoto and Tetsuya Shigeyasu

Enhancement of Quality Assurance Controls in a Smart Transportation System: Application to Petrol Product Distribution 204
 Rexhina Hoxha, Eva Mandri, Artemisa Sinorukaj, Elinda Kajo Meçe, Roberto Sacile, Ilir Shinko, and Enrico Zero

Hardware-Software Interworking Real-Time V2X Dynamic Analysis Method 215
 Insu Oh, Munkhdelgerekh Batzorig, Baasantogtokh Duulga, and Kangbin Yim

Location-Based Autonomous Transmission Control Method for Spatio-Temporal Data Retention System 224
 Daiki Nobayashi, Kazuya Tsukamoto, Takeshi Ikenaga, and Myung Lee

Vehicle Routing in Whole and Segmented Areas to Incrementally Collect the Disaster Information 236
 Sanjukta Khwairakpam, Masahiro Shibata, and Masato Tsuru

Towards a Methodology for the Semantic Representation of Iot Sensors and BPMNs to Discover Business Process Patterns: A Smart Irrigation Case Study 248
 Beniamino Di Martino, Luigi Colucci Cante, Antonio Esposito, and Mariangela Graziano

Applying CI/CD Process to Improve the Speed and Critical Quality of Perfective Maintenance 259
 Sen-Tarng Lai and Fang-Yie Leu

Generating Personalized Phishing Emails for Social Engineering Training Based on Neural Language Models 270
 Shih-Wei Guo, Tzu-Chi Chen, Hui-Juan Wang, Fang-Yie Leu, and Yao-Chung Fan

Stock Price Trend Prediction Using LSTM and Sentiment Analysis on News Headlines 282
 Jung-Bin Li, Szu-Yin Lin, Fang-Yie Leu, and Yen-Chu Chu

Efficient Weighted and Balanced Resource Allocation for High-Performance Render Farms 292
 Lung-Pin Chen, Fang-Yie Leu, Chia-Chen Kuo, Tzu-Ching Lin, and Ming-Jen Wang

**A Brake Assisting Function for Railway Vehicles Using Fuzzy Logic:
A Comparison Study for Different Fuzzy Inference Types** 301
Mitsuki Tsuneyoshi, Makoto Ikeda, and Leonard Barolli

Preliminary Analysis of Performance Variation for ADS-B Position . . . 312
Junichi Honda, Keisuke Matsunaga, Yasuyuki Kakubari,
and Takuya Otsuyama

**A Simulation System for Mobility Control of Swarm Drones
to Provide Wireless Mesh Network Services.** 323
Yuma Yamashita, Nobuki Saito, Chihiro Yukawa, Kyohei Toyoshima,
Tetsuya Oda, Kengo Katayama, and Leonard Barolli

**Comparison of Transmission Spectra of Fork-Shaped Photonic
Crystal Branch Waveguide for Continuous and Band-Limited
Input Signal** 332
Hiroshi Maeda

Design and Implementation of a Platform for MOAP Robots 341
Keita Matsuo, Elis Kulla, and Leonard Barolli

**Design of an Intelligent Robotic Vision System for Optimization
of Robot Arm Movement** 353
Chihiro Yukawa, Nobuki Saito, Aoto Hirata, Kyohei Toyoshima,
Yuki Nagai, Tetsuya Oda, and Leonard Barolli

**A Transportation Routing Method Based on A* Algorithm and Hill
Climbing for Swarm Robots in WLAN Environment.** 361
Masahiro Niihara, Nobuki Saito, Chihiro Yukawa, Kyohei Toyoshima,
Tetsuya Oda, Masaharu Hirota, and Leonard Barolli

Simulation of Choice of Residence for Working Women 369
Risa Takata, Shiori Koga, and Kaoru Fujioka

**Constructing and Reconstructing Characters Using Gaussian
Process Regression** 377
Jinya Yano and Hiroyuki Fujioka

**Proposal of Disaster Prevention Training System Using Mixed
Reality Space** 385
Takahiro Uchiya and Kazuki Akita

Author Index. 395



A Comparison Study of UNDX and UNDX-m Methods for LDVM and RDVM Router Replacement Methods by WMN-PSODGA Hybrid Intelligent System Considering Stadium Distribution

Admir Barolli¹(✉), Kevin Bylykbashi², Ermioni Qafzezi³, Shinji Sakamoto⁴,
Leonard Barolli², and Makoto Takizawa⁵

¹ Department of Information Technology, Aleksander Moisiu University of Durres,
L.1, Rruga e Currilave, Durres, Albania

admirbarolli@uamd.edu.al

² Department of Information and Communication Engineering, Fukuoka Institute
of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan

kevin@bene.fit.ac.jp, barolli@fit.ac.jp

³ Graduate School of Engineering, Fukuoka Institute of Technology,
3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan

bd20101@bene.fit.ac.jp

⁴ Department of Information and Computer Science, Kanazawa Institute
of Technology, 7-1 Ohgigaoka, Nonoichi, Ishikawa 921-8501, Japan

shinji.sakamoto@ieee.org

⁵ Department of Advanced Sciences, Faculty of Science and Engineering,
Hosei University, 3-7-2, Kajino-machi, Koganei-shi, Tokyo 184-8584, Japan

makoto.takizawa@computer.org

Abstract. Wireless Mesh Networks (WMNs) are gaining a lot of attention from researchers due to their advantages such as easy maintenance, low upfront cost, and high robustness. However, designing a robust WMN at low cost requires the use of the least possible mesh routers but still interconnected and able to offer full coverage. Therefore, the placement of mesh routers over the area of interest is a problem that entails thorough planning. In our previous work, we implemented a simulation system that deals with this problem considering Particle Swarm Optimization (PSO) and Distributed Genetic Algorithm (DGA), called WMN-PSODGA. In this paper, we compare the results of Stadium distribution of mesh clients for Unimodal Normal Distribution Crossover (UNDX) and Multi-parental UNDX (UNDX-m) methods for two router replacement methods: Linearly Decreasing Vmax Method (LDVM) and Rational Decrement of Vmax Method (RDVM). The simulation results show that the use of UNDX with RDVM achieves full client coverage, better connectivity and improved load balance.

1 Introduction

The wireless networks and devices are becoming increasingly popular and they provide users access to information and communication anytime and anywhere [2, 8, 11, 19]. Wireless Mesh Networks (WMNs) are gaining a lot of attention because of their low-cost nature that makes them attractive for providing wireless Internet connectivity. A WMN is dynamically self-organized and self-configured, with the nodes in the network automatically establishing and maintaining mesh connectivity among themselves (creating, in effect, an ad hoc network). This feature brings many advantages to WMN such as low up-front cost, easy network maintenance, robustness and reliable service coverage [1]. Moreover, such infrastructure can be used to deploy community networks, metropolitan area networks, municipal and corporate networks, and to support applications for urban areas, medical, transport and surveillance systems.

Mesh node placement in WMNs can be seen as a family of problems, which is shown (through graph theoretic approaches or placement problems, e.g., [6, 12]) to be computationally hard to solve for most of the formulations [23].

We consider the version of the mesh router nodes placement problem in which we are given a grid area where to deploy a number of mesh router nodes and a number of mesh client nodes of fixed positions (of an arbitrary distribution) in the grid area. The objective is to assign the mesh routers such locations in the grid area that achieve the maximum network connectivity and client coverage while balancing the load among mesh routers. Network connectivity is measured by Size of Giant Component (SGC) of the resulting WMN graph, while the user coverage is simply the number of mesh client nodes that fall within the radio coverage of at least one mesh router node and is measured by the Number of Covered Mesh Clients (NCMC). For load balancing, we added in the fitness function a new parameter called NCMCpR (Number of Covered Mesh Clients per Router).

Node placement problems are known to be computationally hard to solve [9, 10, 24]. In previous works, some intelligent algorithms have been recently investigated for the node placement problem [3, 7, 13, 15].

In [17], we implemented a Particle Swarm Optimization (PSO) based simulation system, called WMN-PSO. Also, we implemented another simulation system based on Genetic Algorithms (GA), called WMN-GA [16], for solving the node placement problem in WMNs. Then, we designed and implemented a hybrid simulation system based on PSO and Distributed GA (DGA). We call this system WMN-PSODGA.

In [4], we implemented two crossover methods: Unimodal Normal Distribution Crossover (UNDX) and Multi-parental UNDX (UNDX-m), and compared their results considering the Stadium distribution of mesh clients and Rational Decrement of Vmax Method (RDVM) as a router replacement method. In this paper, we implement the crossover methods with Linearly Decreasing Vmax Method (LDVM) and compare the results achieved for each combination of crossover and router replacement methods.

The rest of the paper is organized as follows. In Sect. 2, we introduce intelligent algorithms. In Sect. 3 is presented the implemented hybrid simulation system. The simulation results are given in Sect. 4. Finally, we give conclusions and future work in Sect. 5.

2 Intelligent Algorithms for Proposed Hybrid Simulation System

2.1 Particle Swarm Optimization

In PSO a number of simple entities (the particles) are placed in the search space of some problem or function whereby each particle evaluates the objective function at its current location. The objective function is often minimized and the exploration of the search space is not performed through evolution [14].

Each particle then determines its movement through the search space by combining some aspect of the history of its own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have updated their solutions. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function.

Each individual in the particle swarm is composed of three \mathcal{D} -dimensional vectors, where \mathcal{D} is the dimensionality of the search space. These are the current position \vec{x}_i , the previous best position \vec{p}_i and the velocity \vec{v}_i .

The particle swarm is more than just a collection of particles. A particle by itself has almost no power to solve any problem; progress occurs only when the particles interact with each other. Problem solving is a population-wide phenomenon, emerging from the individual behaviors of the particles through their interactions. In any case, populations are organized according to some sort of communication structure or topology, often thought of as a social network. The topology typically consists of bidirectional edges connecting pairs of particles, so that if j is in i 's neighborhood, i is also in j 's. Each particle communicates with some other particles and is affected by the best point found by any member of its topological neighborhood. This is just the vector \vec{p}_i for that best neighbor, which we will denote with \vec{p}_g . The potential kinds of population "social networks" are hugely varied, but in practice certain types have been used more frequently. We show the pseudo code of PSO in Algorithm 1.

In the PSO process, the velocity of each particle is iteratively adjusted so that the particle stochastically oscillates around \vec{p}_i and \vec{p}_g locations.

Algorithm 1. Pseudo code of PSO.

```

/* Initialize all parameters for PSO */
Computation maxtime:=  $Tp_{max}$ ,  $t := 0$ ;
Number of particle-patterns:=  $m$ ,  $2 \leq m \in \mathbf{N}^1$ ;
Particle-patterns initial solution:=  $\mathbf{P}_i^0$ ;
Particle-patterns initial position:=  $\mathbf{x}_{ij}^0$ ;
Particles initial velocity:=  $\mathbf{v}_{ij}^0$ ;
PSO parameter:=  $\omega$ ,  $0 < \omega \in \mathbf{R}^1$ ;
PSO parameter:=  $C_1$ ,  $0 < C_1 \in \mathbf{R}^1$ ;
PSO parameter:=  $C_2$ ,  $0 < C_2 \in \mathbf{R}^1$ ;
/* Start PSO */
Evaluate( $\mathbf{G}^0, \mathbf{P}^0$ );
while  $t < Tp_{max}$  do
    /* Update velocities and positions */
     $\mathbf{v}_{ij}^{t+1} = \omega \cdot \mathbf{v}_{ij}^t$ 
         $+ C_1 \cdot \text{rand}() \cdot (\text{best}(P_{ij}^t) - \mathbf{x}_{ij}^t)$ 
         $+ C_2 \cdot \text{rand}() \cdot (\text{best}(G^t) - \mathbf{x}_{ij}^t)$ ;
     $\mathbf{x}_{ij}^{t+1} = \mathbf{x}_{ij}^t + \mathbf{v}_{ij}^{t+1}$ ;
    /* if fitness value is increased, a new solution will be accepted. */
    Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
     $t = t + 1$ ;
end while
Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
return Best found pattern of particles as solution;

```

2.2 Distributed Genetic Algorithm

Distributed Genetic Algorithm (DGA) has been used in various fields of science. DGA has shown their usefulness for the resolution of many computationally hard combinatorial optimization problems. We show the pseudo code of DGA in Algorithm 2.

Population of individuals: Unlike local search techniques that construct a path in the solution space jumping from one solution to another through local perturbations, DGA use a population of individuals, giving thus the search a larger scope and chances to find better solutions. This feature is also known as the “exploration” process in difference to the “exploitation” process of local search methods.

Fitness: The determination of an appropriate fitness function, together with the chromosome encoding are crucial to the performance of DGA. Ideally we would construct objective functions with “certain regularities”, i.e. objective functions that verify that for any two individuals which are close in the search space, their respective values in the objective functions are similar.

Selection: The selection of individuals to be crossed is another important aspect in DGA as it impacts on the convergence of the algorithm. Several selection schemes have been proposed in the literature for selection operators trying to cope with premature convergence of DGA. There are many selection methods

Algorithm 2. Pseudo code of DGA.

```

/* Initialize all parameters for DGA */
Computation maxtime:=  $Tg_{max}$ ,  $t := 0$ ;
Number of islands:=  $n$ ,  $1 \leq n \in \mathbf{N}^1$ ;
initial solution:=  $\mathbf{P}_i^0$ ;
/* Start DGA */
Evaluate( $\mathbf{G}^0, \mathbf{P}^0$ );
while  $t < Tg_{max}$  do
    for all islands do
        Selection();
        Crossover();
        Mutation();
    end for
     $t = t + 1$ ;
end while
Update.Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
return Best found pattern of particles as solution;

```

in GA. In our system, we implement 2 selection methods: the random method and the roulette wheel method.

Crossover operators: Use of crossover operators is one of the most important characteristics. Crossover operator is the means of DGA to transmit best genetic features of parents to offsprings during generations of the evolution process. Many methods for crossover operators have been proposed such as Unimodal Normal Distribution Crossover (UNDX), Multi-parental UNDX (UNDX-m), Simplex Crossover (SPX), and Blend Crossover (BLX- α). In this paper, we implement and compare the results of the first two methods.

Mutation operators: These operators intend to improve the individuals of a population by small local perturbations. They aim to provide a component of randomness in the neighborhood of the individuals of the population. In our system, we implemented two mutation methods: uniformly random mutation and boundary mutation.

Escaping from local optima: GA itself has the ability to avoid falling prematurely into local optima and can eventually escape from them during the search process. DGA has one more mechanism to escape from local optima by considering some islands. Each island computes GA for optimizing and they migrate their genes to provide the ability of avoiding local optima (see Fig. 1).

Convergence: The convergence of the algorithm is the mechanism of DGA to reach to good solutions. A premature convergence of the algorithm would cause that all individuals of the population be similar in their genetic features and thus the search would result ineffective and the algorithm getting stuck into local optima. Maintaining the diversity of the population is therefore very important to this family of evolutionary algorithms.

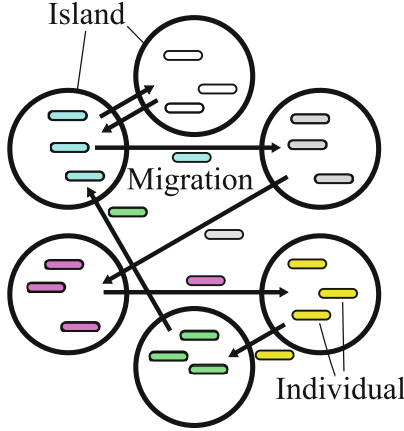


Fig. 1. Model of migration in DGA.

Algorithm 3. Pseudo code of WMN-PSODGA system.

```

Computation maxtime:=  $T_{max}$ ,  $t := 0$ ;
Initial solutions:  $\mathbf{P}$ .
Initial global solutions:  $\mathbf{G}$ .
/* Start PSODGA */
while  $t < T_{max}$  do
  Subprocess(PSO);
  Subprocess(DGA);
  WaitSubprocesses();
  Evaluate( $\mathbf{G}^t, \mathbf{P}^t$ )
  /* Migration() swaps solutions (see Fig. 2). */
  Migration();
   $t = t + 1$ ;
end while
Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
return Best found pattern of particles as solution;

```

3 Proposed and Implemented WMN-PSODGA Hybrid Intelligent Simulation System

In this section, we present the proposed WMN-PSODGA hybrid intelligent simulation system. In the following, we describe the initialization, particle-pattern, gene coding, fitness function, and replacement methods. The pseudo code of our implemented system is shown in Algorithm 3. Also, our implemented simulation system uses Migration function as shown in Fig. 2. The Migration function swaps solutions among lands included in PSO part.

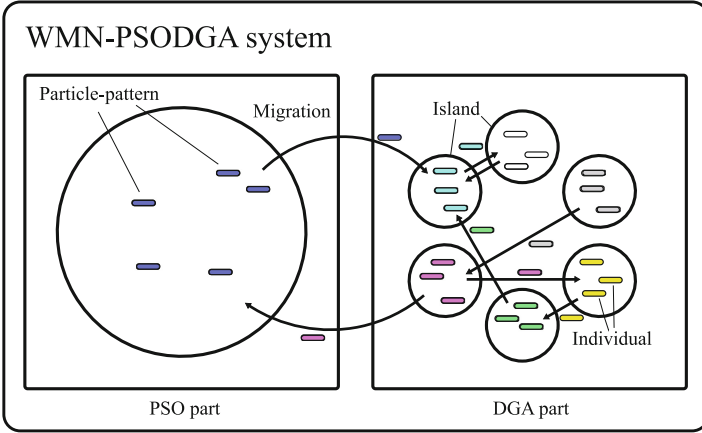


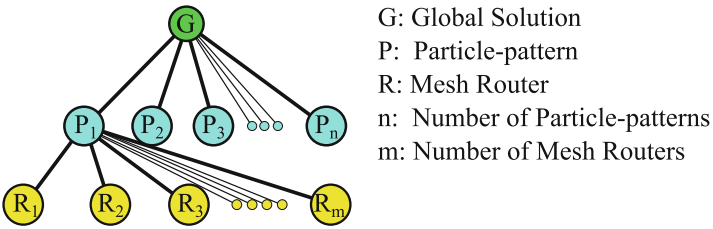
Fig. 2. Model of WMN-PSODGA migration.

Initialization

We decide the velocity of particles by a random process considering the area size. For instance, when the area size is $W \times H$, the velocity is decided randomly from $-\sqrt{W^2 + H^2}$ to $\sqrt{W^2 + H^2}$.

Particle-Pattern

A particle is a mesh router. A fitness value of a particle-pattern is computed by combination of mesh routers and mesh clients positions. In other words, each particle-pattern is a solution as shown is Fig. 3.



G: Global Solution
 P: Particle-pattern
 R: Mesh Router
 n: Number of Particle-patterns
 m: Number of Mesh Routers

Fig. 3. Relationship among global solution, particle-patterns, and mesh routers in PSO part.

Gene Coding

A gene describes a WMN. Each individual has its own combination of mesh nodes. In other words, each individual has a fitness value. Therefore, the combination of mesh nodes is a solution.

Fitness Function

WMN-PSODGA has the fitness function to evaluate the temporary solution of the routers' placements. The fitness function is defined as:

$$Fitness = \alpha \times NCMC(\mathbf{x}_{ij}, \mathbf{y}_{ij}) + \beta \times SGC(\mathbf{x}_{ij}, \mathbf{y}_{ij}) + \gamma \times NCMCpR(\mathbf{x}_{ij}, \mathbf{y}_{ij}).$$

This function uses the following indicators.

- NCMC (Number of Covered Mesh Clients)
The NCMC is the number of the clients covered by the SGC's routers.
- SGC (Size of Giant Component)
The SGC is the maximum number of connected routers.
- NCMCpR (Number of Covered Mesh Clients per Router)
The NCMCpR is the number of clients covered by each router. The NCMCpR indicator is used for load balancing.

WMN-PSODGA aims to maximize the value of the fitness function in order to optimize the placement of the routers using the above three indicators. Weight-coefficients of the fitness function are α , β , and γ for NCMC, SGC, and NCMCpR, respectively. Moreover, the weight-coefficients are implemented as $\alpha + \beta + \gamma = 1$.

Router Replacement Methods

A mesh router has x , y positions, and velocity. Mesh routers are moved based on velocities. There are many router replacement methods. In this paper, we consider the Linearly Decreasing Vmax Method (LDVM) and Rational Decrement of Vmax Method (RDVM).

Constriction Method (CM)

CM is a method in which the PSO parameters are set to a week stable region ($\omega = 0.729$, $C_1 = C_2 = 1.4955$) based on analysis of PSO by M. Clerc et al. [5, 21].

Random Inertia Weight Method (RIWM)

In RIWM, the ω parameter changes randomly from 0.5 to 1.0. The C_1 and C_2 are kept 2.0. The ω can be estimated by the week stable region. The average of ω is 0.75 [21].

Linearly Decreasing Inertia Weight Method (LDIWM)

In LDIWM, C_1 and C_2 are set to 2.0, constantly. On the other hand, the ω parameter changes linearly from the unstable region ($\omega = 0.9$) to the stable region ($\omega = 0.4$) after each iteration [21, 22].

Linearly Decreasing Vmax Method (LDVM)

In LDVM, PSO parameters are set to the unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). A value of V_{max} , which is the maximum velocity of the particles, is considered. The V_{max} decreases linearly after each iteration [20].

Table 1. The common parameters for each simulation.

Parameters	Values
Distribution of mesh clients	Stadium
Number of mesh clients	48
Number of mesh routers	17
Radius of a mesh router	2.0–3.5
Number of GA islands	16
Number of migrations	200
Evolution steps	9
Selection method	Random
Crossover method	UNDX, UNDX-m
Mutation method	Uniform
Crossover rate	0.8
Mutation rate	0.2
Replacement method	LDVM, RDVM
Area size	32.0×32.0

Rational Decrement of Vmax Method (RDVM)

In RDVM, PSO parameters are set to the unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). The V_{max} decreases after each iteration as below:

$$V_{max}(x) = \sqrt{W^2 + H^2} \times \frac{T - x}{x}.$$

Where W and H are the width and the height of the considered area, respectively. Also, T and x are the total number of iterations and the current number of iterations, respectively [18].

4 Simulation Results

In this section, we present and compare the simulation results of UNDX and UNDX-m considering Stadium distribution of mesh clients for LDVM and RDVM router replacement methods. The weight-coefficients of the fitness function were adjusted for optimization. In this paper, the weight-coefficients are $\alpha = 0.8$, $\beta = 0.1$, $\gamma = 0.1$. The number of mesh routers and mesh clients is 17 and 48, whereas the selection and mutation methods are Random and Uniform, respectively. Table 1 summarizes the common parameters used for the simulations. Figure 4 and Fig. 5 show the visualization results after the optimization for LDVM and RDVM, respectively. Figure 6 and Fig. 7 show the number of covered mesh clients by each router, whereas Fig. 8 and 9 the standard deviation where r is the correlation coefficient.

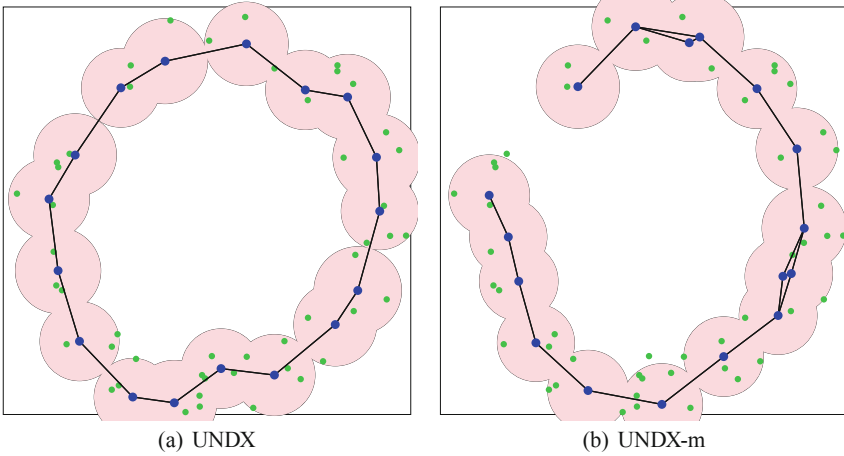


Fig. 4. Visualization results after the optimization [RDVM].

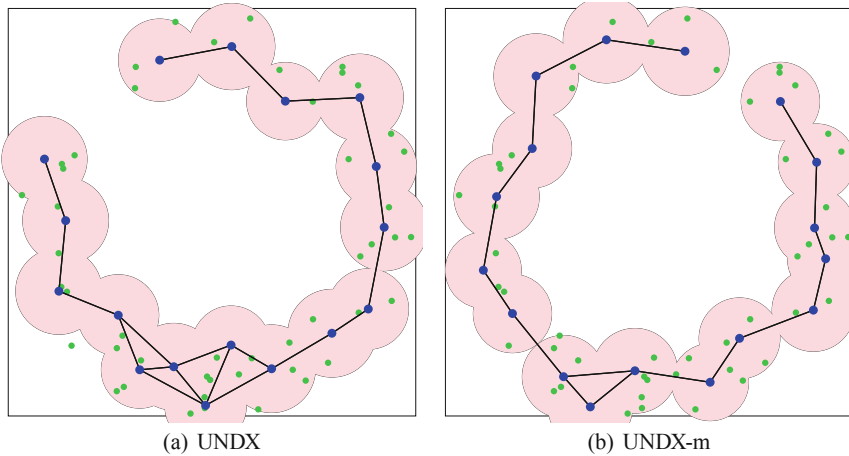


Fig. 5. Visualization results after the optimization [LDVM].

As shown in Fig. 4, the simulation results show that better coverage and connectivity is achieved when UNDX is used. When using UNDX-m the mesh routers do not cover all mesh clients, however the routers are all connected to each other. On the other hand, the simulation results show that when using LDVM as a router replacement method (see Fig. 5), UNDX-m outperforms UNDX as it achieves full coverage while still connecting all routers to each other.

In Fig. 6 and Fig. 7 we see that in each simulation scenario each mesh router covers at least two mesh clients and at most seven mesh clients. These results indicate that every combination router replacement-crossover method achieves good balancing. However, Fig. 8 and Fig. 9 show which of the combinations

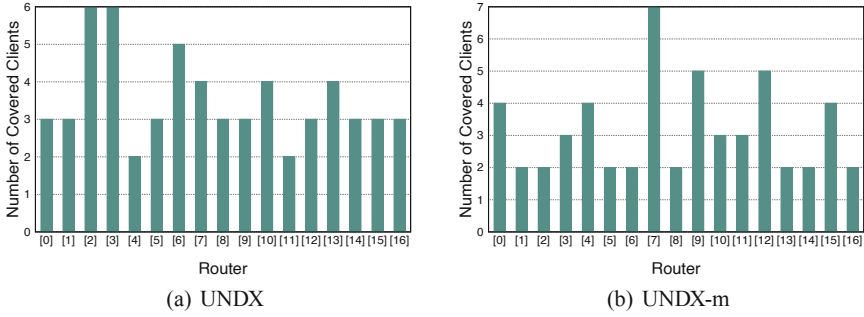


Fig. 6. Number of covered clients by each router after the optimization [RDVM].

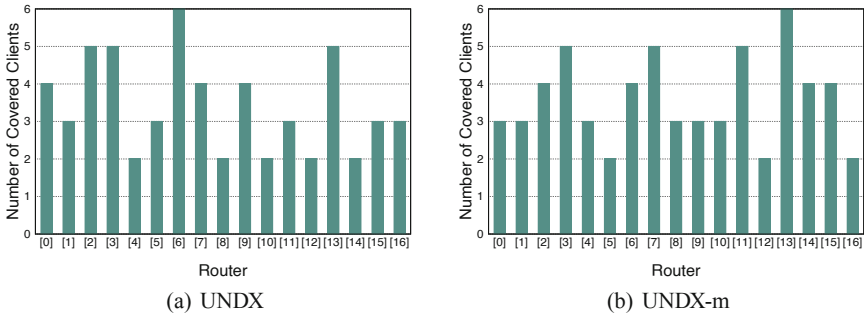


Fig. 7. Number of covered clients by each router after the optimization [LDVM].

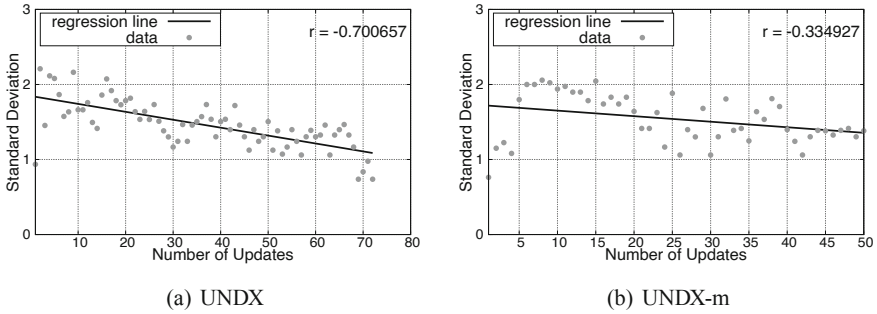


Fig. 8. Transition of the standard deviations [RDVM].

achieves the best results by means of comparing their standard deviations and their correlation coefficients. When the standard deviation is an increasing line ($r > 0$), the number of mesh clients for each router tends to be different. On the other hand, when the standard deviation is a decreasing line ($r < 0$), the number of mesh clients for each router tends to go close to each other. The standard deviation is a decreasing line in each case, but a better load balancing is achieved when RDVM and UNDX are used together.

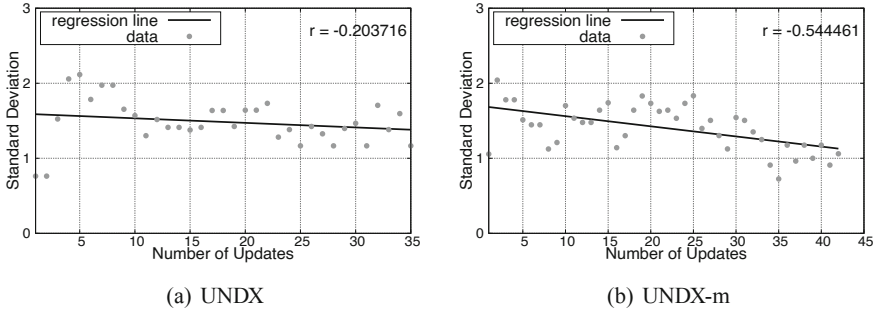


Fig. 9. Transition of the standard deviations [LDVM].

5 Conclusions

In this work, we evaluated the performance of WMNs using a hybrid simulation system based on PSO and DGA (called WMN-PSODGA). We compared the simulation results of RDVM and LDVM router replacement methods for the Stadium distribution of mesh clients using UNDX and UNDX-m as crossover methods.

The simulation results show that for RDVM, UNDX achieves full client coverage, better connectivity and improved load balance. Acceptable connectivity and load balancing are achieved when using UNDX-m, too, but not all mesh clients are covered in this case. On the other hand, when using LDVM as a router replacement method, UNDX-m outperforms UNDX, but it does not achieve the connectivity and load balancing of the UNDX-RDVM combination.

In future work, we will consider the implementation of crossover methods with different mutation and other router replacement methods.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. *Comput. Netw.* **47**(4), 445–487 (2005)
2. Barolli, A., Sakamoto, S., Barolli, L., Takizawa, M.: Performance analysis of simulation system based on particle swarm optimization and distributed genetic algorithm for WMNs considering different distributions of mesh clients. In: Barolli, L., Khafa, F., Javaid, N., Enokido, T. (eds.) *IMIS 2018. AISC*, vol. 773, pp. 32–45. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-93554-6_3
3. Barolli, A., Sakamoto, S., Ozera, K., Barolli, L., Kulla, E., Takizawa, M.: Design and implementation of a hybrid intelligent system based on particle swarm optimization and distributed genetic algorithm. In: Barolli, L., Khafa, F., Javaid, N., Spaho, E., Kolic, V. (eds.) *EIDWT 2018. LNDECT*, vol. 17, pp. 79–93. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75928-9_7
4. Barolli, A., Bylykbashi, K., Qafzezi, E., Sakamoto, S., Barolli, L., Takizawa, M.: Mesh routers placement by WMN-PSODGA simulation system considering stadium distribution and RDVM: a comparison study for UNDX and UNDX-M

- methods. In: Barolli, L. (ed.) IMIS 2022. Lecture Notes in Networks and Systems, vol. 496, pp. 184–195. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08819-3_18
5. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evol. Comput.* **6**(1), 58–73 (2002)
 6. Franklin, A.A., Murthy, C.S.R.: Node placement algorithm for deployment of two-tier wireless mesh networks. In: Proceedings of Global Telecommunications Conference, pp. 4823–4827 (2007)
 7. Girgis, M.R., Mahmoud, T.M., Abdullatif, B.A., Rabie, A.M.: Solving the wireless mesh network design problem using genetic algorithm and simulated annealing optimization methods. *Int. J. Comput. Appl.* **96**(11), 1–10 (2014)
 8. Goto, K., Sasaki, Y., Hara, T., Nishio, S.: Data gathering using mobile agents for reducing traffic in dense mobile wireless sensor networks. *Mob. Inf. Syst.* **9**(4), 295–314 (2013)
 9. Lim, A., Rodrigues, B., Wang, F., Xu, Z.: k-center problems with minimum coverage. *Theoret. Comput. Sci.* **332**(1–3), 1–17 (2005)
 10. Maolin, T., et al.: Gateways placement in backbone wireless mesh networks. *Int. J. Commun. Netw. Syst. Sci.* **2**(1), 44–50 (2009)
 11. Matsuo, K., Sakamoto, S., Oda, T., Barolli, A., Ikeda, M., Barolli, L.: Performance analysis of WMNs by WMN-GA simulation system for two WMN architectures and different TCP congestion-avoidance algorithms and client distributions. *Int. J. Commun. Netw. Distrib. Syst.* **20**(3), 335–351 (2018)
 12. Muthaiah, S.N., Rosenberg, C.P.: Single gateway placement in wireless mesh networks. In: Proceedings of 8th International IEEE Symposium on Computer Networks, pp. 4754–4759 (2008)
 13. Naka, S., Genji, T., Yura, T., Fukuyama, Y.: A hybrid particle swarm optimization for distribution state estimation. *IEEE Trans. Power Syst.* **18**(1), 60–68 (2003)
 14. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. *Swarm Intell.* **1**(1), 33–57 (2007)
 15. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A comparison study of simulated annealing and genetic algorithm for node placement problem in wireless mesh networks. *J. Mob. Multimed.* **9**(1–2), 101–110 (2013)
 16. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A comparison study of hill climbing, simulated annealing and genetic algorithm for node placement problem in WMNs. *J. High Speed Netw.* **20**(1), 55–66 (2014)
 17. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation and evaluation of a simulation system based on particle swarm optimisation for node placement problem in wireless mesh networks. *Int. J. Commun. Netw. Distrib. Syst.* **17**(1), 1–13 (2016)
 18. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation of a new replacement method in WMN-PSO simulation system and its performance evaluation. In: The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016), pp. 206–211 (2016)
 19. Sakamoto, S., Ozera, K., Ikeda, M., Barolli, L.: Implementation of intelligent hybrid systems for node placement problem in WMNs considering particle swarm optimization, hill climbing and simulated annealing. *Mob. Netw. Appl.* **23**(1), 27–33 (2017). <https://doi.org/10.1007/s11036-017-0897-7>
 20. Schutte, J.F., Groenwold, A.A.: A study of global optimization using particle swarms. *J. Global Optim.* **31**(1), 93–108 (2005)
 21. Shi, Y.: Particle swarm optimization. *IEEE Connect.* **2**(1), 8–13 (2004)

22. Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E. (eds.) EP 1998. LNCS, vol. 1447, pp. 591–600. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0040810>
23. Vanhatupa, T., Hannikainen, M., Hamalainen, T.: Genetic algorithm to optimize node placement and configuration for WLAN planning. In: Proceedings of the 4th IEEE International Symposium on Wireless Communication Systems, pp. 612–616 (2007)
24. Wang, J., Xie, B., Cai, K., Agrawal, D.P.: Efficient mesh router placement in wireless mesh networks. In: Proceedings of IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS-2007), pp. 1–9 (2007)



Performance Comparison of Roulette Wheel and Random Selection Methods by WMN-PSODGA Simulation System Considering Stadium Distribution and LDIWM

Kevin Bylykbashi¹✉, Ermioni Qafzezi², Phudit Ampririt², Admir Barolli³, Elis Kulla⁴, and Leonard Barolli¹

¹ Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
kevin@bene.fit.ac.jp, barolli@fit.ac.jp

² Graduate School of Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
{bd20101,bd20201}@bene.fit.ac.jp

³ Department of Information Technology, Aleksander Moisiu University of Durres, L.1, Rruga e Currilave, Durres, Albania

⁴ Department of System Management, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
kulla@fit.ac.jp

Abstract. Wireless Mesh Networks (WMNs) are gaining a lot of attention from researchers due to their advantages such as easy maintenance, low upfront cost, and high robustness. However, designing a robust WMN at low cost requires the use of the least possible mesh routers but still interconnected and able to offer full coverage. Therefore, the placement of mesh routers over the area of interest is a problem that entails thorough planning. In our previous work, we implemented a simulation system that deals with this problem considering Particle Swarm Optimization (PSO) and Distributed Genetic Algorithm (DGA), called WMN-PSODGA. In this paper, we compare the results of Stadium distribution of mesh clients for Roulette Wheel and Random Selection methods for Linearly Decreasing Inertia Weight Method (LDIWM). The simulation results show that the use of roulette wheel achieves full client coverage, better connectivity and improved load balance.

1 Introduction

The wireless networks and devices are becoming increasingly popular and they provide users access to information and communication anytime and anywhere [2, 7, 10]. Wireless Mesh Networks (WMNs) are gaining a lot of attention because of their low-cost nature that makes them attractive for providing wireless Internet connectivity. A WMN is dynamically self-organized and self-configured,

with the nodes in the network automatically establishing and maintaining mesh connectivity among themselves (creating, in effect, an ad hoc network). This feature brings many advantages to WMN such as low up-front cost, easy network maintenance, robustness and reliable service coverage [1]. Moreover, such infrastructure can be used to deploy community networks, metropolitan area networks, municipal and corporate networks, and to support applications for urban areas, medical, transport and surveillance systems.

Mesh node placement in WMNs can be seen as a family of problems, which is shown (through graph theoretic approaches or placement problems, e.g., [5, 11]) to be computationally hard to solve for most of the formulations [21].

We consider the version of the mesh router nodes placement problem in which we are given a grid area where to deploy a number of mesh router nodes and a number of mesh client nodes of fixed positions (of an arbitrary distribution) in the grid area. The objective is to assign the mesh routers such locations in the grid area that achieve the maximum network connectivity and client coverage while balancing the load among mesh routers. Network connectivity is measured by Size of Giant Component (SGC) of the resulting WMN graph, while the user coverage is simply the number of mesh client nodes that fall within the radio coverage of at least one mesh router node and is measured by the Number of Covered Mesh Clients (NCMC). For load balancing, we added in the fitness function a new parameter called NCMCpR (Number of Covered Mesh Clients per Router).

Node placement problems are known to be computationally hard to solve [8, 9, 22]. In previous works, some intelligent algorithms have been recently investigated for the node placement problem [3, 6, 12, 14].

In [16], we implemented a Particle Swarm Optimization (PSO) based simulation system, called WMN-PSO. Also, we implemented another simulation system based on Genetic Algorithms (GA), called WMN-GA [15], for solving the node placement problem in WMNs. Then, we designed and implemented a hybrid simulation system based on PSO and Distributed GA (DGA). We call this system WMN-PSODGA.

In this paper, we implement the roulette wheel and random selection methods and compare the results that these two selection methods achieve for the Stadium distribution of mesh clients when using Linearly Decreasing Inertia Weight Method (LDIWM) as a router replacement method.

The rest of the paper is organized as follows. In Sect. 2, we introduce intelligent algorithms. In Sect. 3 is presented the implemented hybrid simulation system. The simulation results are given in Sect. 4. Finally, we give conclusions and future work in Sect. 5.

2 Intelligent Algorithms for Proposed Hybrid Simulation System

2.1 Particle Swarm Optimization

In PSO a number of simple entities (the particles) are placed in the search space of some problem or function whereby each particle evaluates the objective function at its current location. The objective function is often minimized and the exploration of the search space is not performed through evolution [13].

Each particle then determines its movement through the search space by combining some aspect of the history of its own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have updated their solutions. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function.

Each individual in the particle swarm is composed of three \mathcal{D} -dimensional vectors, where \mathcal{D} is the dimensionality of the search space. These are the current position \vec{x}_i , the previous best position \vec{p}_i and the velocity \vec{v}_i .

The particle swarm is more than just a collection of particles. A particle by itself has almost no power to solve any problem; progress occurs only when the particles interact with each other. Problem solving is a population-wide phenomenon, emerging from the individual behaviors of the particles through their interactions. In any case, populations are organized according to some sort of communication structure or topology, often thought of as a social network. The topology typically consists of bidirectional edges connecting pairs of particles, so that if j is in i 's neighborhood, i is also in j 's. Each particle communicates with some other particles and is affected by the best point found by any member of its topological neighborhood. This is just the vector \vec{p}_i for that best neighbor, which we will denote with \vec{p}_g . The potential kinds of population “social networks” are hugely varied, but in practice certain types have been used more frequently. We show the pseudo code of PSO in Algorithm 1.

In the PSO process, the velocity of each particle is iteratively adjusted so that the particle stochastically oscillates around \vec{p}_i and \vec{p}_g locations.

2.2 Distributed Genetic Algorithm

Distributed Genetic Algorithm (DGA) has been used in various fields of science. DGA has shown their usefulness for the resolution of many computationally hard combinatorial optimization problems. We show the pseudo code of DGA in Algorithm 2.

Population of individuals: Unlike local search techniques that construct a path in the solution space jumping from one solution to another through local perturbations, DGA use a population of individuals, giving thus the search a larger scope and chances to find better solutions. This feature is also known as the “exploration” process in difference to the “exploitation” process of local search methods.

Algorithm 1. Pseudo code of PSO.

```

/* Initialize all parameters for PSO */
Computation maxtime:=  $Tp_{max}$ ,  $t := 0$ ;
Number of particle-patterns:=  $m$ ,  $2 \leq m \in \mathbf{N}^1$ ;
Particle-patterns initial solution:=  $\mathbf{P}_i^0$ ;
Particle-patterns initial position:=  $\mathbf{x}_{ij}^0$ ;
Particles initial velocity:=  $\mathbf{v}_{ij}^0$ ;
PSO parameter:=  $\omega$ ,  $0 < \omega \in \mathbf{R}^1$ ;
PSO parameter:=  $C_1$ ,  $0 < C_1 \in \mathbf{R}^1$ ;
PSO parameter:=  $C_2$ ,  $0 < C_2 \in \mathbf{R}^1$ ;
/* Start PSO */
Evaluate( $\mathbf{G}^0, \mathbf{P}^0$ );
while  $t < Tp_{max}$  do
  /* Update velocities and positions */
   $\mathbf{v}_{ij}^{t+1} = \omega \cdot \mathbf{v}_{ij}^t$ 
     $+ C_1 \cdot \text{rand}() \cdot (\text{best}(P_{ij}^t) - \mathbf{x}_{ij}^t)$ 
     $+ C_2 \cdot \text{rand}() \cdot (\text{best}(G^t) - \mathbf{x}_{ij}^t)$ ;
   $\mathbf{x}_{ij}^{t+1} = \mathbf{x}_{ij}^t + \mathbf{v}_{ij}^{t+1}$ ;
  /* if fitness value is increased, a new solution will be accepted. */
  Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
   $t = t + 1$ ;
end while
Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
return Best found pattern of particles as solution;

```

Fitness: The determination of an appropriate fitness function, together with the chromosome encoding are crucial to the performance of DGA. Ideally we would construct objective functions with “certain regularities”, i.e. objective functions that verify that for any two individuals which are close in the search space, their respective values in the objective functions are similar.

Selection: The selection of individuals to be crossed is another important aspect in DGA as it impacts on the convergence of the algorithm. Several selection schemes have been proposed in the literature for selection operators trying to cope with premature convergence of DGA. There are many selection methods in GA. In our system, we implement 2 selection methods: the random method and the roulette wheel method.

Crossover operators: Use of crossover operators is one of the most important characteristics. Crossover operator is the means of DGA to transmit best genetic features of parents to offsprings during generations of the evolution process. Many methods for crossover operators have been proposed such as Unimodal Normal Distribution Crossover (UNDX), Multi-parental UNDX (UNDX-m), Simplex Crossover (SPX), and Blend Crossover (BLX- α). In this paper, we use UNDX.

Mutation operators: These operators intend to improve the individuals of a population by small local perturbations. They aim to provide a component of randomness in the neighborhood of the individuals of the population. In our

Algorithm 2. Pseudo code of DGA.

```

/* Initialize all parameters for DGA */
Computation maxtime:=  $Tg_{max}$ ,  $t := 0$ ;
Number of islands:=  $n$ ,  $1 \leq n \in \mathbf{N}^1$ ;
initial solution:=  $\mathbf{P}_i^0$ ;
/* Start DGA */
Evaluate( $\mathbf{G}^0, \mathbf{P}^0$ );
while  $t < Tg_{max}$  do
    for all islands do
        Selection();
        Crossover();
        Mutation();
    end for
     $t = t + 1$ ;
end while
Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
return Best found pattern of particles as solution;

```

system, we implemented two mutation methods: uniformly random mutation and boundary mutation.

Escaping from local optima: GA itself has the ability to avoid falling prematurely into local optima and can eventually escape from them during the search process. DGA has one more mechanism to escape from local optima by considering some islands. Each island computes GA for optimizing and they migrate their genes to provide the ability of avoiding local optima (see Fig. 1).

Convergence: The convergence of the algorithm is the mechanism of DGA to reach to good solutions. A premature convergence of the algorithm would cause that all individuals of the population be similar in their genetic features and thus the search would result ineffective and the algorithm getting stuck into local optima. Maintaining the diversity of the population is therefore very important to this family of evolutionary algorithms.

3 Proposed and Implemented WMN-PSODGA Hybrid Intelligent Simulation System

In this section, we present the proposed WMN-PSODGA hybrid intelligent simulation system. In the following, we describe the initialization, particle-pattern, gene coding, fitness function, and replacement methods. The pseudo code of our implemented system is shown in Algorithm 3. Also, our implemented simulation system uses Migration function as shown in Fig. 2. The Migration function swaps solutions among lands included in PSO part.

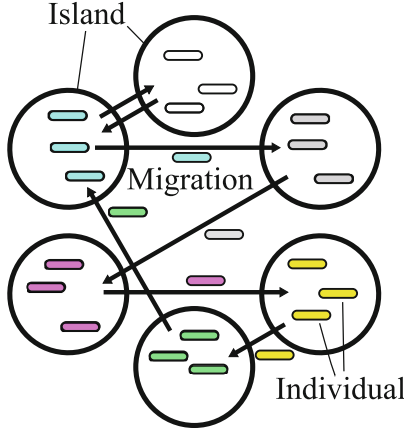


Fig. 1. Model of migration in DGA.

Algorithm 3. Pseudo code of WMN-PSODGA system.

```

Computation maxtime:=  $T_{max}$ ,  $t := 0$ ;
Initial solutions:  $\mathbf{P}$ .
Initial global solutions:  $\mathbf{G}$ .
/* Start PSODGA */
while  $t < T_{max}$  do
  Subprocess(PSO);
  Subprocess(DGA);
  WaitSubprocesses();
  Evaluate( $\mathbf{G}^t, \mathbf{P}^t$ )
  /* Migration() swaps solutions (see Fig. 2). */
  Migration();
   $t = t + 1$ ;
end while
Update_Solutions( $\mathbf{G}^t, \mathbf{P}^t$ );
return Best found pattern of particles as solution;

```

Initialization

We decide the velocity of particles by a random process considering the area size. For instance, when the area size is $W \times H$, the velocity is decided randomly from $-\sqrt{W^2 + H^2}$ to $\sqrt{W^2 + H^2}$.

Particle-pattern

A particle is a mesh router. A fitness value of a particle-pattern is computed by combination of mesh routers and mesh clients positions. In other words, each particle-pattern is a solution as shown in Fig. 3.

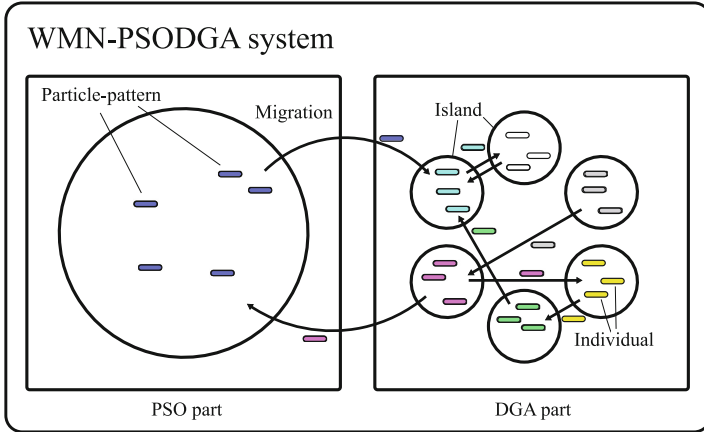
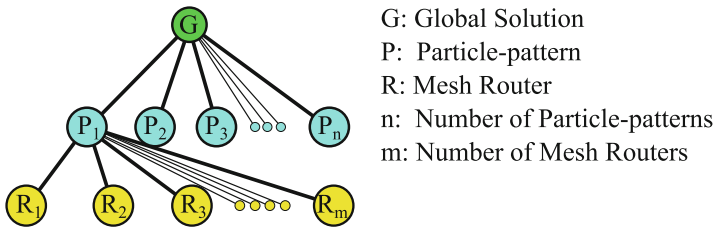


Fig. 2. Model of WMN-PSODGA migration.



G: Global Solution
 P: Particle-pattern
 R: Mesh Router
 n: Number of Particle-patterns
 m: Number of Mesh Routers

Fig. 3. Relationship among global solution, particle-patterns, and mesh routers in PSO part.

Gene Coding

A gene describes a WMN. Each individual has its own combination of mesh nodes. In other words, each individual has a fitness value. Therefore, the combination of mesh nodes is a solution.

Fitness Function

WMN-PSODGA has the fitness function to evaluate the temporary solution of the routers’ placements. The fitness function is defined as:

$$Fitness = \alpha \times NCMC(\mathbf{x}_{ij}, \mathbf{y}_{ij}) + \beta \times SGC(\mathbf{x}_{ij}, \mathbf{y}_{ij}) + \gamma \times NCMCpR(\mathbf{x}_{ij}, \mathbf{y}_{ij}).$$

This function uses the following indicators.

- NCMC (Number of Covered Mesh Clients)
 The NCMC is the number of the clients covered by the SGC’s routers.

- SGC (Size of Giant Component)
The SGC is the maximum number of connected routers.
- NCMCpR (Number of Covered Mesh Clients per Router)
The NCMCpR is the number of clients covered by each router. The NCMCpR indicator is used for load balancing.

WMN-PSODGA aims to maximize the value of the fitness function in order to optimize the placement of the routers using the above three indicators. Weight-coefficients of the fitness function are α , β , and γ for NCMC, SGC, and NCMCpR, respectively. Moreover, the weight-coefficients are implemented as $\alpha + \beta + \gamma = 1$.

Router Replacement Methods

A mesh router has x , y positions, and velocity. Mesh routers are moved based on velocities. There are many router replacement methods. In this paper, we consider the Linearly Decreasing Inertia Weight Method (LDIWM).

Constriction Method (CM)

CM is a method in which the PSO parameters are set to a week stable region ($\omega = 0.729$, $C_1 = C_2 = 1.4955$) based on analysis of PSO by M. Clerc et al. [4, 19].

Random Inertia Weight Method (RIWM)

In RIWM, the ω parameter changes randomly from 0.5 to 1.0. The C_1 and C_2 are kept 2.0. The ω can be estimated by the week stable region. The average of ω is 0.75 [19].

Linearly Decreasing Inertia Weight Method (LDIWM)

In LDIWM, C_1 and C_2 are set to 2.0, constantly. On the other hand, the ω parameter changes linearly from the unstable region ($\omega = 0.9$) to the stable region ($\omega = 0.4$) after each iteration [19, 20].

Linearly Decreasing Vmax Method (LDVM)

In LDVM, PSO parameters are set to the unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). A value of V_{max} , which is the maximum velocity of the particles, is considered. The V_{max} decreases linearly after each iteration [18].

Rational Decrement of Vmax Method (RDVM)

In RDVM, PSO parameters are set to the unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). The V_{max} decreases after each iteration as below:

$$V_{max}(x) = \sqrt{W^2 + H^2} \times \frac{T - x}{x}.$$

Where W and H are the width and the height of the considered area, respectively. Also, T and x are the total number of iterations and the current number of iterations, respectively [17].

Table 1. The common parameters for each simulation.

Parameters	Values
Distribution of mesh clients	Stadium
Number of mesh clients	48
Number of mesh routers	16
Radius of a mesh router	2.0–3.5
Number of GA islands	16
Number of migrations	200
Evolution steps	9
Selection method	Roulette wheel, random
Crossover method	UNDX
Mutation method	Uniform
Crossover rate	0.8
Mutation rate	0.2
Replacement method	LDIWM
Area size	32.0×32.0

4 Simulation Results

In this section, we present and compare the simulation results of roulette wheel and random selection methods considering Stadium distribution of mesh clients and the LDIWM router replacement method. The weight-coefficients of the fitness function were adjusted for optimization. In this paper, the weight-coefficients are $\alpha = 0.8$, $\beta = 0.1$, $\gamma = 0.1$. The number of mesh routers and mesh clients is 16 and 48, whereas the crossover and mutation methods are UNDX and Uniform, respectively. Table 1 summarizes the common parameters used for the simulations. Figure 4 shows the coverage results after the optimization, Fig. 5 the number of covered mesh clients by each router, whereas Fig. 6 the standard deviation where r is the correlation coefficient.

As shown in Fig. 4, the simulation results show that better coverage and connectivity is achieved when the roulette wheel method is used. When using the random method the mesh routers do not cover all mesh clients; however, all routers are connected to each other.

In Fig. 5, we see that in each simulation scenario each mesh router covers at least two mesh clients and at most seven mesh clients. These results indicate both methods achieve good load balancing. Figure 6 shows which of the combinations achieves the best results by means of comparing their standard deviations and correlation coefficients. When the standard deviation is an increasing line ($r > 0$), the number of mesh clients for each router tends to be different. On the other hand, when the standard deviation is a decreasing line ($r < 0$), the number of mesh clients for each router tends to go close to each other. The standard deviation is a decreasing line in each case, but a better load balancing is achieved when roulette wheel is used.

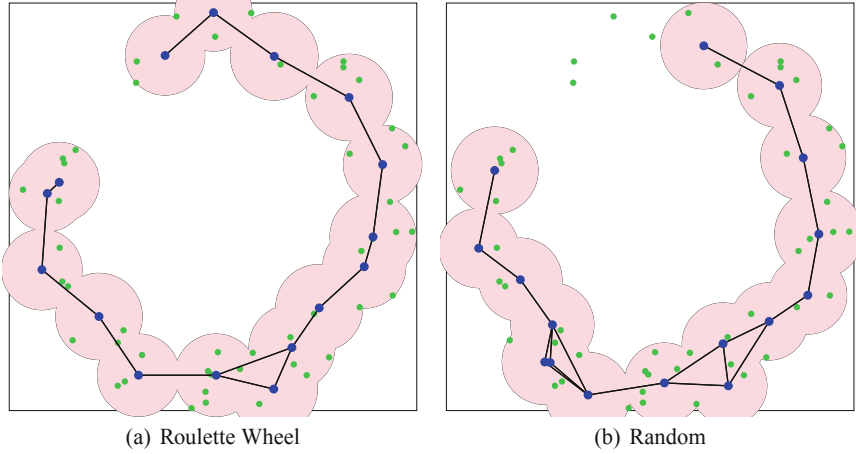


Fig. 4. Visualization results after the optimization.

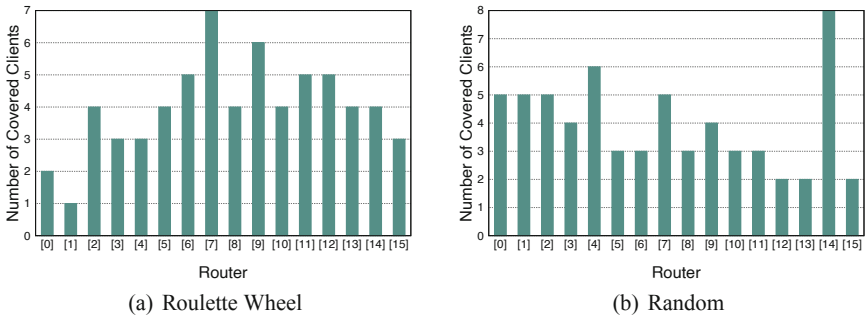


Fig. 5. Number of covered clients by each router after the optimization.

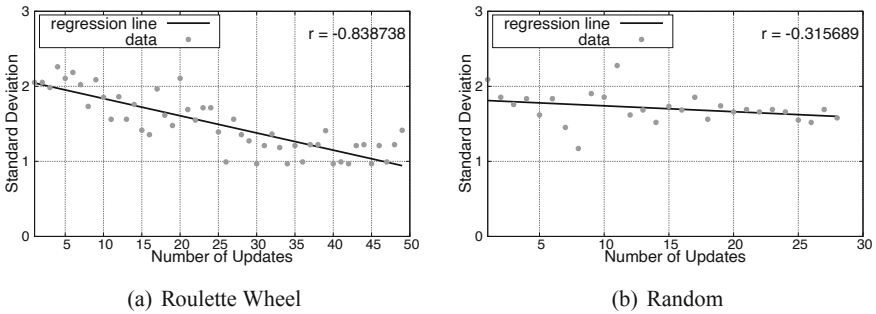


Fig. 6. Transition of the standard deviations.

5 Conclusions

In this work, we evaluated the performance of WMNs using a hybrid simulation system based on PSO and DGA (called WMN-PSODGA). We compared the simulation results of roulette wheel and random selection methods for the Stadium distribution of mesh clients using LDIWM as a router replacement method.

The simulation results show that the roulette wheel selection method achieves full client coverage, better connectivity and improved load balance. Acceptable connectivity and load balancing are achieved when using the random method, too, but not all mesh clients are covered in this case.

In future work, we will consider the combination of selection methods with different mutation and other router replacement methods.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. *Comput. Netw.* **47**(4), 445–487 (2005)
2. Barolli, A., Sakamoto, S., Barolli, L., Takizawa, M.: Performance analysis of simulation system based on particle swarm optimization and distributed genetic algorithm for WMNs considering different distributions of mesh clients. In: Barolli, L., Xhafa, F., Javaid, N., Enokido, T. (eds.) *IMIS 2018. AISC*, vol. 773, pp. 32–45. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-93554-6_3
3. Barolli, A., Sakamoto, S., Ozera, K., Barolli, L., Kulla, E., Takizawa, M.: Design and implementation of a hybrid intelligent system based on particle swarm optimization and distributed genetic algorithm. In: Barolli, L., Xhafa, F., Javaid, N., Spaho, E., Kolici, V. (eds.) *EIDWT 2018. LNDECT*, vol. 17, pp. 79–93. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75928-9_7
4. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evol. Comput.* **6**(1), 58–73 (2002)
5. Franklin, A.A., Murthy, C.S.R.: Node placement algorithm for deployment of two-tier wireless mesh networks. In: *Proceedings of Global Telecommunications Conference*, pp. 4823–4827 (2007)
6. Girgis, M.R., Mahmoud, T.M., Abdullatif, B.A., Rabie, A.M.: Solving the wireless mesh network design problem using genetic algorithm and simulated annealing optimization methods. *Int. J. Comput. Appl.* **96**(11), 1–10 (2014)
7. Goto, K., Sasaki, Y., Hara, T., Nishio, S.: Data gathering using mobile agents for reducing traffic in dense mobile wireless sensor networks. *Mob. Inf. Syst.* **9**(4), 295–314 (2013)
8. Lim, A., Rodrigues, B., Wang, F., Xu, Z.: k-Center problems with minimum coverage. *Theoret. Comput. Sci.* **332**(1–3), 1–17 (2005)
9. Maolin, T., et al.: Gateways placement in backbone wireless mesh networks. *Int. J. Commun. Netw. Syst. Sci.* **2**(1), 44–50 (2009)
10. Matsuo, K., Sakamoto, S., Oda, T., Barolli, A., Ikeda, M., Barolli, L.: Performance analysis of WMNs by WMN-GA simulation system for two WMN architectures and different TCP congestion-avoidance algorithms and client distributions. *Int. J. Commun. Netw. Distrib. Syst.* **20**(3), 335–351 (2018)
11. Muthaiah, S.N., Rosenberg, C.P.: Single gateway placement in wireless mesh networks. In: *Proceedings of 8th International IEEE Symposium on Computer Networks*, pp. 4754–4759 (2008)

12. Naka, S., Genji, T., Yura, T., Fukuyama, Y.: A hybrid particle swarm optimization for distribution state estimation. *IEEE Trans. Power Syst.* **18**(1), 60–68 (2003)
13. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. *Swarm Intell.* **1**(1), 33–57 (2007). <https://doi.org/10.1007/s11721-007-0002-0>
14. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A comparison study of simulated annealing and genetic algorithm for node placement problem in wireless mesh networks. *J. Mob. Multimed.* **9**(1–2), 101–110 (2013)
15. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A comparison study of hill climbing, simulated annealing and genetic algorithm for node placement problem in WMNs. *J. High Speed Netw.* **20**(1), 55–66 (2014)
16. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation and evaluation of a simulation system based on particle swarm optimisation for node placement problem in wireless mesh networks. *Int. J. Commun. Netw. Distrib. Syst.* **17**(1), 1–13 (2016)
17. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation of a new replacement method in WMN-PSO simulation system and its performance evaluation. In: *The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016)*, pp 206–211 (2016)
18. Schutte, J.F., Groenwold, A.A.: A study of global optimization using particle swarms. *J. Glob. Optim.* **31**(1), 93–108 (2005)
19. Shi, Y.: Particle swarm optimization. *IEEE Connect.* **2**(1), 8–13 (2004)
20. Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E. (eds.) *EP 1998. LNCS*, vol. 1447, pp. 591–600. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0040810>
21. Vanhatupa, T., Hannikainen, M., Hamalainen, T.: Genetic algorithm to optimize node placement and configuration for WLAN planning. In: *Proceedings of the 4th IEEE International Symposium on Wireless Communication Systems*, pp. 612–616 (2007)
22. Wang, J., Xie, B., Cai, K., Agrawal, D.P.: Efficient mesh router placement in wireless mesh networks. In: *Proceedings of IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS-2007)*, pp. 1–9 (2007)



A Fuzzy-Based System for Handover in 5G Wireless Networks Considering Different Network Slicing Constraints: Effects of Slice Reliability Parameter on Handover Decision

Phudit Ampririt¹(✉), Ermioni Qafzezi¹, Kevin Bylykbashi², Makoto Ikeda², Keita Matsuo², and Leonard Barolli²

¹ Graduate School of Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
{bd21201,bd20101}@bene.fit.ac.jp

² Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan

kevin@bene.fit.ac.jp, makoto.ikd@acm.org, {kt-matsuo,barolli}@fit.ac.jp

Abstract. Handover in 5G wireless networks introduces new and complex challenges because a user not only changes between base stations or access technologies, but also between slices. Users should select the slices that satisfy their requirements or preferences. When making a handover decision to satisfy user requirements, the constraints on Network Slicing (NS) should be considered. In this paper, we propose a Fuzzy-based system for Handover considering four parameters: Slice Delay (SD), Slice Bandwidth (SB), Slice Stability (SS) and Slice Reliability (SR) as a new parameter. From simulation results, we conclude that the considered parameters have different effects on the Handover Decision (HD). When SD is increased, the HD parameter is increased. But when SB, SS and SR are increasing, the HD parameter is decreased.

1 Introduction

In 5G wireless networks, the massive growth of users device with unpredictable traffic patterns will create large data volume on the Internet, causing congestion and Quality of Service (QoS) deterioration [1]. Also, Handover process is a critical component for mobility management and can affect the overall network performance [2]. Many Handover (HO) scenarios and different HO rates may occur, which bring problems on ensuring stable and reliable connections [3].

For dealing with these problems, the 5G Wireless Networks will provide increased performance in terms of throughput, latency, reliability and mobility in order to fulfill the QoS requirements in many application scenarios. The 5G is developing by considering three main different usage scenarios which have been identified as enhanced mobile broadband (eMBB), ultra-reliable & low latency communications (URLLC) and massive type communication (mMTC). The eMBB is related to human-essential and has greater accessibility to multimedia content and services by enhancing seamless Quality of Experience (QoE). The URLLC can efficiently reduce the latency and enhance reliability. The mMTC can accommodate a large number of connected devices while maintaining a long battery life [4–6].

Recently, several research projects are attempting to develop systems that are suited for the 5G era. One of them is the Software-Defined Network (SDN) [7]. Also, introducing Fuzzy Logic (FL) to SDN controllers enhance the QoS. In addition, the mobile handover method with SDN is used to reduce the processing delays [8–10].

In our previous work [11, 12], we presented a Fuzzy-based system for Handover in 5G Wireless Networks considering four input parameters: Slice Delay (SD), Slice Bandwidth (SB), Slice Stability (SS) and Slice Load (SL). The output parameter was Handover Decision (HD). In this paper, we propose a Fuzzy-based system for Handover in 5G Wireless Networks considering four parameters: SD, SB, SS and Slice Reliability (SR) as a new parameter.

The rest of the paper is organized as follows. In Sect. 2 is presented an overview of SDN. In Sect. 3, we present 5G Network Slicing. In Sect. 4, we describe the proposed Fuzzy-based system and its implementation. In Sect. 5, we discuss the simulation results. Finally, conclusions and future work are presented in Sect. 6.

2 Software-Defined Networks (SDNs)

The SDN is a new networking paradigm that decouples the data plane from control plane in the network. By SDN is easy to manage and provide network software based services from a centralised control plane. The SDN control plane is managed by SDN controller or cooperating group of SDN controllers. The SDN structure is shown in Fig. 1 [13, 14].

- **Application Layer** builds an abstracted view of the network by collecting information from the controller for decision-making purposes. The types of applications are related to network configuration and management, network monitoring, network troubleshooting, network policies and security.
- **Northbound Interfaces** allow communication between the control layer and the application layer and can provide a lot of possibilities for networking programming. Based on the needs of the application, it will pass commands and information to the control layer and make the controller creates the best possible software network with suitable qualities of service and acceptable security.

- **Control Layer** receives instructions or requirements from the Application Layer. It contains the controllers that control the data plane and forward the different types of rules and policies to the infrastructure layer through the Southbound Interfaces.
- **Southbound Interfaces** allow connection and interaction between the control plane and the data plane. The southbound interface is defined as protocols that allow the controller to create policies for the forwarding plane.
- **Infrastructure Layer** receives orders from SDN controller and sends data among them. This layer represents the forwarding devices on the network such as routers, switches and load balancers.

The SDN can manage network systems while enabling new services. In congestion traffic situation, the SDN can control and adapt resources appropriately throughout the control plane. Mobility management is easier and quicker in forwarding across different wireless technologies (e.g. 5G, 4G, Wifi and Wimax). Also, the handover procedure is simple and the delay can be decreased.

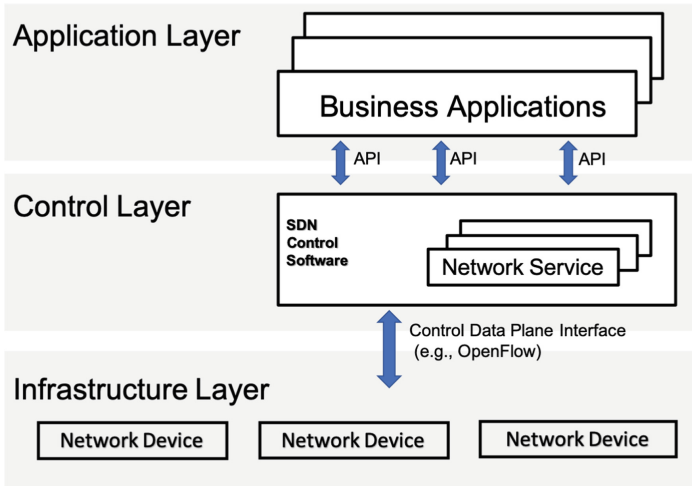


Fig. 1. Structure of SDN.

3 5G Network Slicing

The Network Slicing (NS) is a technology that divides a single virtualized infrastructure into multiple virtual end-to-end networks which are called “Slices”. The NS is configured into virtualized function follow the demand of application to respond to the user’s requests. Each slice is logically independent and doesn’t have any effect on other virtual logical networks [15–17].

A network with NS compared with the traditional networks can provide better performance and can be flexible for different service requirements and number of users. Also, because the slices don't affect each other, the slice reliability and security can be improved [18].

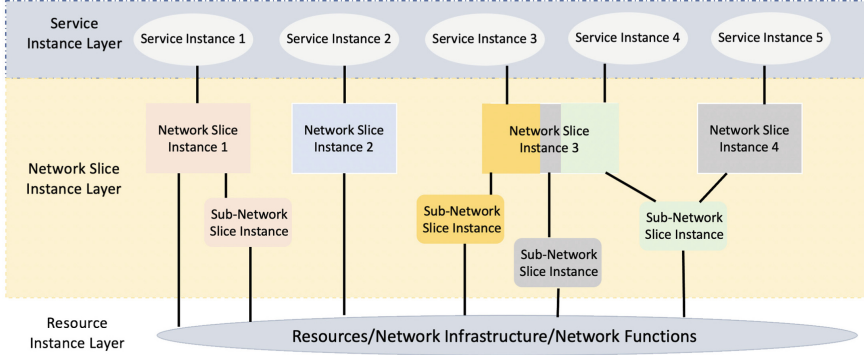


Fig. 2. NGMN NS concept.

The 5G NS concept is developed by the Next Generation Mobile Networks (NGMN) as shown in Fig. 2. The NS process is divided into three main layers [19, 20].

- **Service Instance Layer** represents a service (end-user service or business services) which is provided by application provider or mobile network operator.
- **Network Slice Instance Layer** is a set of network functions and resources which provide the network slice instance to accommodate the required network characteristics (ultra-low-latency, ultra-reliability) by the service instances.
- **Resource Layer** comprises of physical resources and logical resources for the slice deployment.

4 Proposed Fuzzy-Based System

In this work, we use FL to implement the proposed system. In Fig. 3, we show the overview of our proposed system. Each evolve Base Station (eBS) will receive controlling order from the SDN controller and they can communicate and send data with User Equipment (UE). Also, each eBS can cover many slices for different applications. On the other hand, the SDN controller will collect all the data about network traffic status and control eBS to manage inter-eBS handover and inter-slice handover by using the proposed Fuzzy-based system. The SDN controller will be a communication bridge between eBS and the 5G core network.

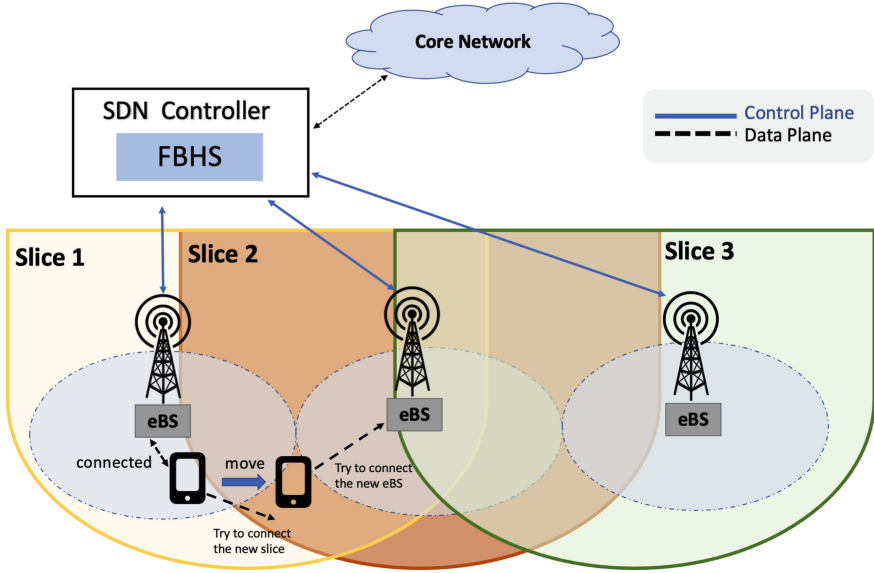


Fig. 3. Proposed system overview.

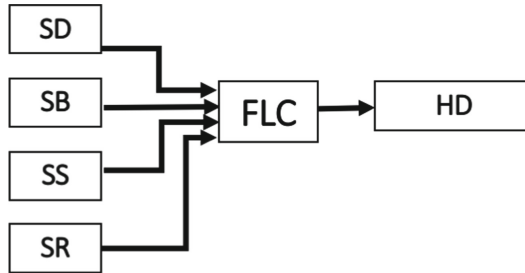


Fig. 4. Proposed system structure.

The proposed system is called Fuzzy-based Handover System (FBHS) in 5G Wireless Networks. The structure of FBHS is shown in Fig. 4. For the implementation of our system, we consider four input parameters: Slice Delay (SD), Slice Bandwidth (SB), Slice Stability (SS), Slice Reliability (SR) as a new parameter and the output parameter is Handover Decision (HD).

Slice Delay (SD): The slice with high delay causes high queuing and link delay. Therefore, the Handover is needed to fulfill the QoS.

Slice Bandwidth (SB): Slice Bandwidth is the available bandwidth of a slice. When SB is higher, the Handover possibility will be lower.

Slice Stability (SS): The slice with high stability can provide consistent communication service and exhibit a stable performance. If the SS is low, the user will consider the handover to another slice with higher stability.

Slice Reliability (SR): When a slice has low reliability, the user will be switched to other slices with higher reliability.

Handover Decision (HD): The HD parameter determines whether or not to perform the handover procedure.

Table 1. Parameter and their term sets.

Parameters	Term sets
Slice Delay (SD)	Low (Lo), Medium (Me), High (Hi)
Slice Bandwidth (SB)	Small (Sm), Intermediate (In), Big (Bi)
Slice Stability (SS)	Low (Lw), Medium (Md), High (Hg)
Slice Reliability (SR)	Low (L), Medium (M), High (H)
Handover Decision (HD)	HD1, HD2, HD3, HD4, HD5, HD6, HD7

The membership functions are shown in Fig. 5. We use triangular and trapezoidal membership functions because they are more suitable for real-time operations [21–24]. We show parameters and their term sets in Table 1. The Fuzzy Rule Base (FRB) is shown in Table 2 and has 81 rules. The control rules have the form: IF “condition” THEN “control action”. For example, for Rule 1: “IF SD is Lo, SB is Sm, SS is Lw and SR is L, THEN HD is HD6”.

5 Simulation Results

In this section, we present the simulation result of our proposed system. The simulation results are shown in Fig. 6, Fig. 7 and Fig. 8. They show the relation of HD with SR for different SS values considering SD and SB as constant parameters.

In Fig. 6, we consider the SD value 0.1 ms. For SB 10% and SS 10%, when SR is increased from 10% to 60% and 60% to 90%, we see that HD is decreased by 15% and 8%, respectively. But, when we increased the SS value from 10% to 90%, the HD value is decreased by 30% when the SR value is 50%. This is because the present slice is more stable and the handover possibility to the other slices is low. When we increased the SB value from 10% to 90%, the HD value is decreased by 30% when the SS value is 10% and the SR value is 60%. This indicates that when the present slice is heavily loaded, the chance of handover to the other slices is high, whereas the chance of handover to the other slices is low when the present slice has more bandwidth.

Table 2. FRB.

Rule	SD	SB	SS	SR	HD	Rule	SD	SB	SS	SR	HD
1	Lo	Sm	Lw	L	HD6	41	Me	In	Md	M	HD4
2	Lo	Sm	Lw	M	HD5	42	Me	In	Md	H	HD3
3	Lo	Sm	Lw	H	HD4	43	Me	In	Hg	L	HD4
4	Lo	Sm	Md	L	HD5	44	Me	In	Hg	M	HD3
5	Lo	Sm	Md	M	HD4	45	Me	In	Hg	H	HD2
6	Lo	Sm	Md	H	HD3	46	Me	Bi	Lw	L	HD5
7	Lo	Sm	Hg	L	HD4	47	Me	Bi	Lw	M	HD4
8	Lo	Sm	Hg	M	HD3	48	Me	Bi	Lw	H	HD3
9	Lo	Sm	Hg	H	HD2	49	Me	Bi	Md	L	HD4
10	Lo	In	Lw	L	HD5	50	Me	Bi	Md	M	HD3
11	Lo	In	Lw	M	HD4	51	Me	Bi	Md	H	HD2
12	Lo	In	Lw	H	HD3	52	Me	Bi	Hg	L	HD3
13	Lo	In	Md	L	HD4	53	Me	Bi	Hg	M	HD2
14	Lo	In	Md	M	HD3	54	Me	Bi	Hg	H	HD1
15	Lo	In	Md	H	HD2	55	Hi	Sm	Lw	L	HD7
16	Lo	In	Hg	L	HD3	56	Hi	Sm	Lw	M	HD7
17	Lo	In	Hg	M	HD2	57	Hi	Sm	Lw	H	HD7
18	Lo	In	Hg	H	HD1	58	Hi	Sm	Md	L	HD7
19	Lo	Bi	Lw	L	HD4	59	Hi	Sm	Md	M	HD7
20	Lo	Bi	Lw	M	HD3	60	Hi	Sm	Md	H	HD6
21	Lo	Bi	Lw	H	HD2	61	Hi	Sm	Hg	L	HD7
22	Lo	Bi	Md	L	HD3	62	Hi	Sm	Hg	M	HD6
23	Lo	Bi	Md	M	HD2	63	Hi	Sm	Hg	H	HD5
24	Lo	Bi	Md	H	HD1	64	Hi	In	Lw	L	HD7
25	Lo	Bi	Hg	L	HD2	65	Hi	In	Lw	M	HD7
26	Lo	Bi	Hg	M	HD1	66	Hi	In	Lw	H	HD6
27	Lo	Bi	Hg	H	HD1	67	Hi	In	Md	L	HD7
28	Me	Sm	Lw	L	HD7	68	Hi	In	Md	M	HD6
29	Me	Sm	Lw	M	HD6	69	Hi	In	Md	H	HD5
30	Me	Sm	Lw	H	HD5	70	Hi	In	Hg	L	HD6
31	Me	Sm	Md	L	HD6	71	Hi	In	Hg	M	HD5
32	Me	Sm	Md	M	HD5	72	Hi	In	Hg	H	HD4
33	Me	Sm	Md	H	HD4	73	Hi	Bi	Lw	L	HD7
34	Me	Sm	Hg	L	HD5	74	Hi	Bi	Lw	M	HD6
35	Me	Sm	Hg	M	HD4	75	Hi	Bi	Lw	H	HD5
36	Me	Sm	Hg	H	HD3	76	Hi	Bi	Md	L	HD6
37	Me	In	Lw	L	HD6	77	Hi	Bi	Md	M	HD5
38	Me	In	Lw	M	HD5	78	Hi	Bi	Md	H	HD4
39	Me	In	Lw	H	HD4	79	Hi	Bi	Hg	L	HD5
40	Me	In	Md	L	HD5	80	Hi	Bi	Hg	M	HD4
						81	Hi	Bi	Hg	H	HD3

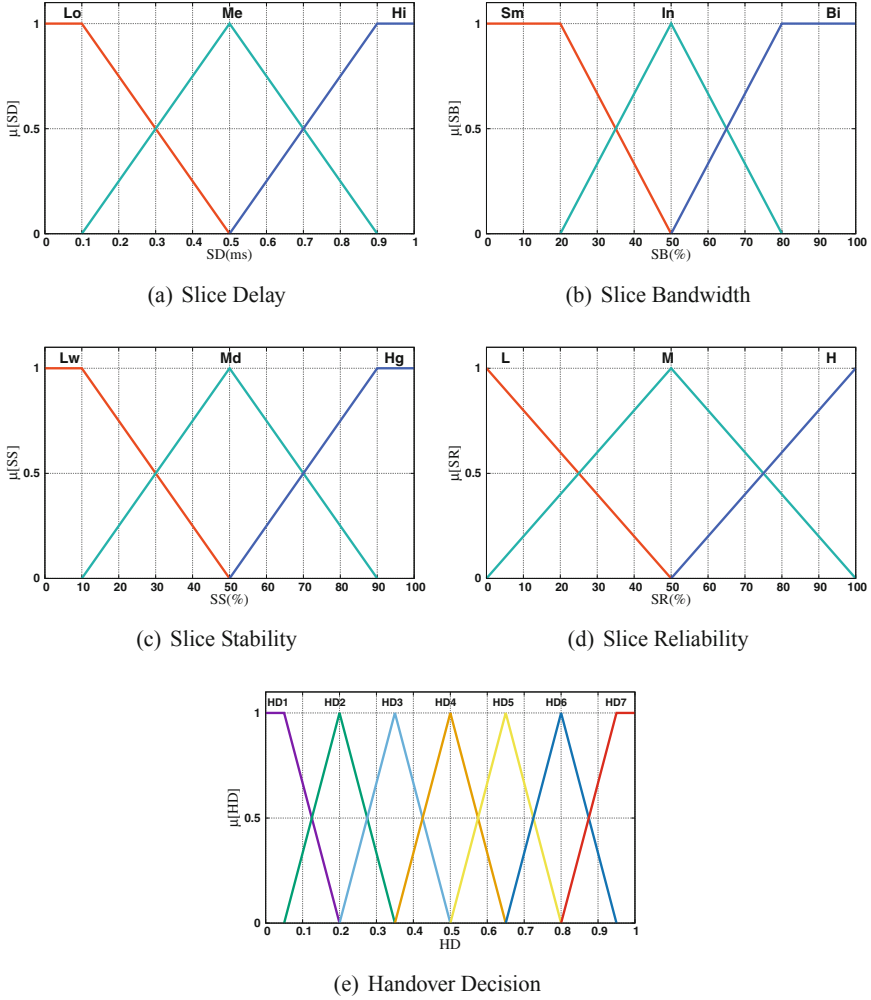


Fig. 5. Membership functions.

We compare Fig. 6 with Fig. 7 to see how SD has affected HD. We change the SD value from 0.1 ms to 0.5 ms. The HD is increased by 15% when the SB value is 10%, the SS is 10% and the SR is 80%. This is because the present slice delay is higher. Thus, the handover to another slice is needed.

We increase the value of SD to 0.9 ms in Fig. 8. Comparing the results with Fig. 6 and Fig. 7, we can see that the HD values have increased significantly. For SD 0.9 ms, SB 10%, all HD values are higher than 0.5. Thus, the mobile device will make a handover to another slice.

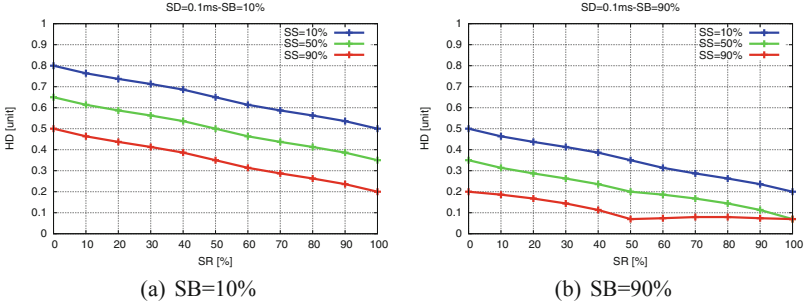


Fig. 6. Simulation results for $SD = 0.1$ ms.

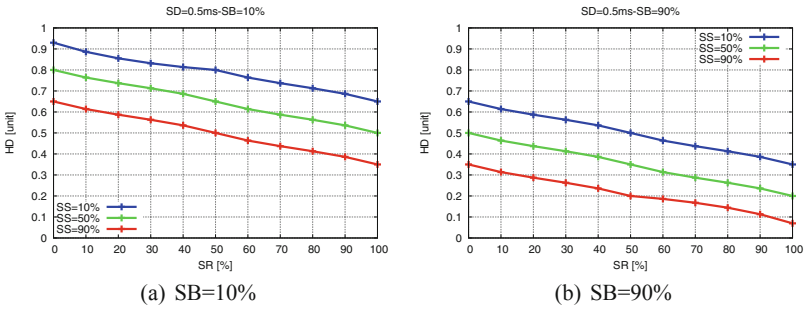


Fig. 7. Simulation results for $SD = 0.5$ ms.

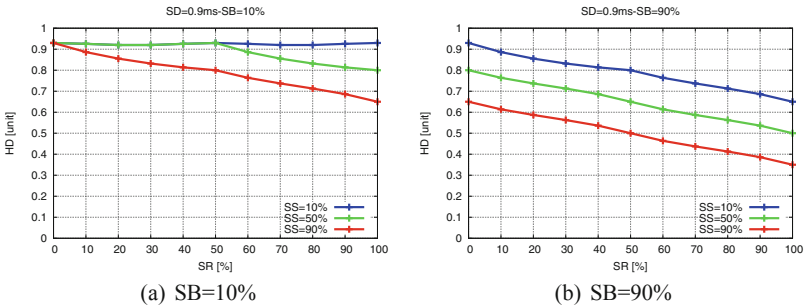


Fig. 8. Simulation results for $SD = 0.9$ ms.

6 Conclusions and Future Work

In this paper, we proposed and implemented a Fuzzy-based system for Handover in 5G Wireless Networks. We considered four parameters: SD, SB, SS and SR to decide the HD value. We evaluated the proposed system by simulations. From the simulation results, we found that four parameters have different effects on the HD. When SD is increased, the HD parameter is increased but when SB, SS and SR are increasing, the HD parameter is decreased.

In the future work, we will consider different parameters and perform extensive simulations to evaluate the proposed system.

References

1. Navarro-Ortiz, J., Romero-Diaz, P., Sendra, S., Ameigeiras, P., Ramos-Munoz, J.J., Lopez-Soler, J.M.: A survey on 5G usage scenarios and traffic models. *IEEE Commun. Surv. Tutorials* **22**(2), 905–929 (2020). <https://doi.org/10.1109/COMST.2020.2971781>
2. Sun, Y., et al.: Efficient handover mechanism for radio access network slicing by exploiting distributed learning. *IEEE Trans. Netw. Serv. Manage.* **17**(4), 2620–2633 (2020). <https://doi.org/10.1109/TNSM.2020.3031079>
3. Saad, W.K., Shayea, I., Hamza, B.J., Mohamad, H., Daradkeh, Y.I., Jabbar, W.A.: Handover parameters optimisation techniques in 5G networks. *Sensors* **21**(15), 5202 (2021). <https://doi.org/10.3390/s21155202>
4. Akpakwu, G.A., Silva, B.J., Hancke, G.P., Abu-Mahfouz, A.M.: A survey on 5G networks for the internet of things: communication technologies and challenges. *IEEE Access* **6**, 3619–3647 (2018)
5. Palmieri, F.: A reliability and latency-aware routing framework for 5G transport infrastructures. *Comput. Netw.* **179**(9), Article 107365 (2020). <https://doi.org/10.1016/j.comnet.2020.107365>
6. Kamil, I.A., Ogundoyin, S.O.: Lightweight privacy-preserving power injection and communication over vehicular networks and 5G smart grid slice with provable security. *Internet Things* **8**(100116), 100–116 (2019). <https://doi.org/10.1016/j.iot.2019.100116>
7. Hossain, E., Hasan, M.: 5G cellular: key enabling technologies and research challenges. *IEEE Instrum. Meas. Mag.* **18**(3), 11–21 (2015). <https://doi.org/10.1109/MIM.2015.7108393>
8. Yao, D., Su, X., Liu, B., Zeng, J.: A mobile handover mechanism based on fuzzy logic and MPTCP protocol under SDN architecture. In: 18th International Symposium on Communications and Information Technologies (ISCIT-2018), pp. 141–146 (2018). <https://doi.org/10.1109/ISCIT.2018.8587956>
9. Lee, J., Yoo, Y.: Handover cell selection using user mobility information in a 5G SDN-based network. In: 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN-2017), pp. 697–702 (2017). <https://doi.org/10.1109/ICUFN.2017.7993880>
10. Moravejosharieh, A., Ahmadi, K., Ahmad, S.: A fuzzy logic approach to increase quality of service in software defined networking. In: 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN-2018), pp. 68–73 (2018). <https://doi.org/10.1109/ICACCCN.2018.8748678>
11. Ampririt, P., Qafzezi, E., Bylykbashi, K., Ikeda, M., Matsuo, K., Barolli, L.: A fuzzy-based system for handover in 5G wireless networks considering network slicing constraints. In: Barolli, L. (ed.) *CISIS 2022*. LNNS, vol. 497, pp. 180–189. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08812-4_18
12. Ampririt, P., Qafzezi, E., Bylykbashi, K., Ikeda, M., Matsuo, K., Barolli, L.: A fuzzy-based system for handover in 5G wireless networks considering different network slicing constraints: effects of slice load parameter on handover decision. In: Barolli, L., Miwa, H., Enokido, T. (eds.) *Advances in Network-Based Information Systems*, pp. 152–162. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-14314-4_15

13. Li, L.E., Mao, Z.M., Rexford, J.: Toward software-defined cellular networks. In: 2012 European Workshop on Software Defined Networking, pp. 7–12 (2012). <https://doi.org/10.1109/EWSDN.2012.28>
14. Mousa, M., Bahaa-Eldin, A.M., Sobh, M.: Software defined networking concepts and challenges. In: 2016 11th International Conference on Computer Engineering & Systems (ICCES-2016), pp. 79–90. IEEE (2016)
15. An, N., Kim, Y., Park, J., Kwon, D.H., Lim, H.: Slice management for quality of service differentiation in wireless network slicing. *Sensors* **19**, 2745 (2019). <https://doi.org/10.3390/s19122745>
16. Jiang, M., Condoluci, M., Mahmoodi, T.: Network slicing management & prioritization in 5G mobile systems. In: European Wireless 2016; 22th European Wireless Conference, pp. 1–6. VDE (2016)
17. Chen, J., et al.: Realizing dynamic network slice resource management based on SDN networks. In: 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp. 120–125 (2019)
18. Li, X., et al.: Network slicing for 5G: challenges and opportunities. *IEEE Internet Comput.* **21**(5), 20–27 (2017)
19. Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A., Flinck, H.: Network slicing and softwarization: a survey on principles, enabling technologies, and solutions. *IEEE Commun. Surv. Tutorials* **20**(3), 2429–2453 (2018). <https://doi.org/10.1109/COMST.2018.2815638>
20. Alliance, N.: Description of network slicing concept. NGMN 5G P **1**(1), 7p (2016). https://ngmn.org/wp-content/uploads/160113-NGMN_Network_Slicing_v1_0.pdf
21. Norp, T.: 5G requirements and key performance indicators. *J. ICT Stand.* **6**(1), 15–30 (2018)
22. Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A.I., Dai, H.: A survey on low latency towards 5G: ran, core network and caching solutions. *IEEE Commun. Surv. Tutorials* **20**(4), 3098–3130 (2018)
23. Kim, Y., Park, J., Kwon, D., Lim, H.: Buffer management of virtualized network slices for quality-of-service satisfaction. In: 2018 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN-2018), pp. 1–4 (2018)
24. Barolli, L., Koyama, A., Yamada, T., Yokoyama, S.: An integrated CAC and routing strategy for high-speed large-scale networks using cooperative agents. *IPSJ J.* **42**(2), 222–233 (2001)



A Simulated Annealing Based Simulation System for Optimization of Wild Deer Damage Prevention Devices

Sora Asada¹, Kyohei Toyoshima², Aoto Hirata², Yuki Nagai², Nobuki Saito², Tetsuya Oda¹(✉), and Leonard Barolli³

¹ Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
{t19j008as,t19j077mt}@ous.jp, oda@ous.ac.jp

² Graduate School of Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama, Okayama 700-0005, Japan
{t22jm24jd,t21jm02zr,t22jm23rv,t21jm01md}@ous.jp

³ Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
barolli@fit.ac.jp

Abstract. Wild deer are damaging agriculture and forestry in Japan. There are many measures such as fences and lights, but none of these measures can prevent wild deer damages. Therefore, wild deer damage prevention measures using ultrasonic waves are attracting attention. The *wild Deer damage prevention Device (DD)* using ultrasonic waves can be a good approach. In this paper, we propose a Simulated Annealing (SA) based *DD* placement optimization system for preventing wild deer damages. In the proposed system, we use the OpenStreetMap (OSM) data of the Ogaya, Nishiwakurason, Okayama Prefecture, Japan as the target to be covered by *DD*. The simulation results show that the proposed system makes a good placement of *DDs* that maximizes network connectivity and the number of *wild Deer damage prevention device Points (DPs)* covered by *DDs*.

1 Introduction

Wild deer are damaging agriculture and forestry in Japan. Fences, lights and other prevention measures are considered to prevent wild deer damages. However, none of these prevention measures are solutions because there are enormous damages. Moreover, conventional prevention measures have several problems such as the cost of annual fence replacement and the effect of light on the nearby residents. Therefore, ultrasonic based wild deer damage prevention measures are studied [1–4].

In Japan, the *wild Deer damage prevention Devices (DDs)* using ultrasonic waves are already available. However, the *DD* are directional and the effect of the prevention measures depends on the *DD* placement location. The placement of

DD also affects the operating costs and operability. On the other hand, effective *DD* placement candidate locations could be decided by using OpenStreetMap (OSM) data for areas that have wild deer damages. In addition, when a *DD* has problem or does not function properly, it is necessary to detect the failure because the wild deer may enter the area. Therefore, we consider that *DDs* can communicate with each other building a wireless sensor actuator network [5–8].

In our previous work [9–12], we proposed and evaluated some node placement optimization methods for Wireless Mesh Networks (WMNs) using Hill Climbing (HC) [13], Simulated Annealing(SA) [14], Genetic Algorithms (GA) [15], and Tabu Search (TS) [16]. We found that SA-based node placement optimization methods covered many mesh clients generated by normal and uniform distributions [17].

In this paper, we propose a SA-based *DD* placement optimization system for preventing wild deer damages. The simulation results show that the proposed system makes a good placement of *DDs* that maximizes network connectivity and the number of *wild Deer damage prevention device Points (DPs)* covered by *DDs*.

The rest of the paper is organized as follows. In Sect. 2, we introduce the proposed system, algorithms. In Sect. 3 is presented the simulation results. Finally, we give conclusions and future work in Sect. 4.

2 Proposed System

2.1 OSM Extraction for *DD* Placement Candidate Locations

The proposed system optimizes the placement of *DDs* in areas that have wild deer damages. The OSM data are used to decide good *DD* placement candidate locations. For the *OSM Extraction of DD placement candidate locations (OSM Extraction)*, we consider OSM data of Ogaya, Nishiawakurason, Okayama Prefecture, Japan. Then, the OSM data are converted to images of only forests and settlements using QGIS.

The color reduction process [18] is performed since color types depend on the resolution in the image output from QGIS. We consider the coordinates of the border between the forest and the settlement as the candidate location of the *DDs* to prevent wild deer from entering the settlement. The image size is scaled down to reduce the processing time of SA based optimization. The *DPs* coordinates are set as the half of the ultrasonic radiation distance of *DDs*, which is the Euclidean distance between the boundary coordinates and the forest coordinates.

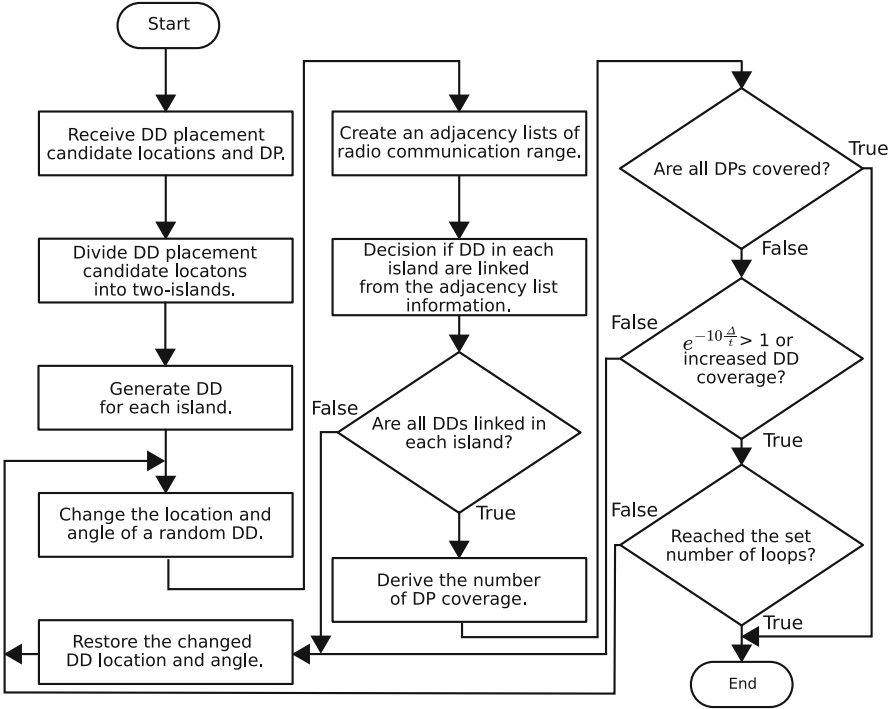


Fig. 1. Flowchart of SA.

2.2 DD Placement Optimization

From *OSM Extraction*, the OSM data of Nishiwakurason are divided into two areas with a road between them. In addition, we consider a two-islands distribution of *DD* placement candidate locations with *DPs* set into an upper and lower area. The upper area is considered as island1 and the lower area as island2. The objective of the *DD* placement optimization is to maximize the network connectivity and the number of *DPs* within the ultrasonic radiation range of *DD* in a continuous two-dimensional region. *DD* placement belongs to the *P-median problem* and is *NP-hard*. The connection graph of the radio communication range of each *DD* is called a *Component* and the *Component* with the large number of connections is called the *Strongly Connected Component (SCC)* [19]. The objective function uses *Size of SCC (SSCC)*, which indicates the network connectivity of the *DD* radio communication range. The *Number of Covered DD (NCDD)* is the number of covered *DPs*.

In two-islands distribution of *DD* placement candidate locations, *SSCC* is used as the objective function because the maximum *SSCC* size in each island maximizes the communication range. There are two *SCCs* in the two-islands distribution of *DD* placement candidate locations. We consider *SCC1* as the upper island and *SCC2* as the bottom island. Also, *NCDD* is the number of *DDs* within the ultrasonic radiation range. *SSCC* is the first objective function and

$NCDD$ is the second objective function to maximize the radio communication range.

The flowchart of the SA-based DD placement optimization system is shown in Fig. 1. First, the DP is divided into an upper island and a lower island. Next, as the initial placement, the DDs are placed at random angles and locations on each island so that the $SSCC$ of each island is maximized. The SA-based process is performed after the initial placement. In SA, the placement of the DDs and the direction of ultrasonic radiation are randomly and repeatedly changed to optimize the placement. The adjacency list is created based on the overlap of radio communication ranges among DDs in each island to derive the SCC . The SCC for each island is derived by Depth-First Search (DFS) [20] performed to the adjacency list. The $NCDD$ is derived from ultrasonic radiation range of DDs covering the DPs after the placement maximizes the SCC of each island.

Table 1. Settings parameter.

Parameter names	Parameter values
$Width \times Height [unit] \times [unit]$	155.0×100.0
Map data scale [$pixel$] : [m]	1 : 0.596
Number of $DD [unit]$	100
Number of DP of $island_1 [unit]$	441
Number of DP of $island_2 [unit]$	396
Radius of communication range [$unit$]	2.5
Ultrasonic radius [$unit$]	10
Ultrasonic angle [$deg.$]	110
Number of iterations [$times$]	25000
Initial temperature [$^{\circ}C$]	100
Final temperature [$^{\circ}C$]	0

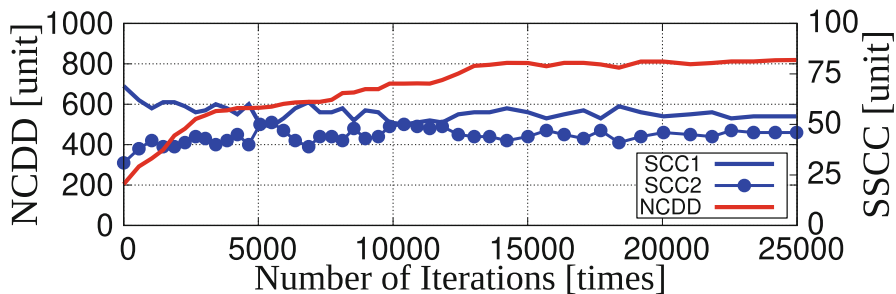
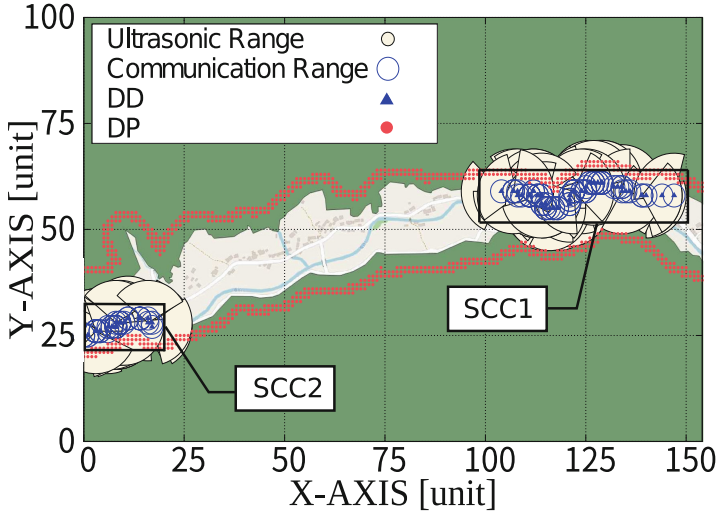
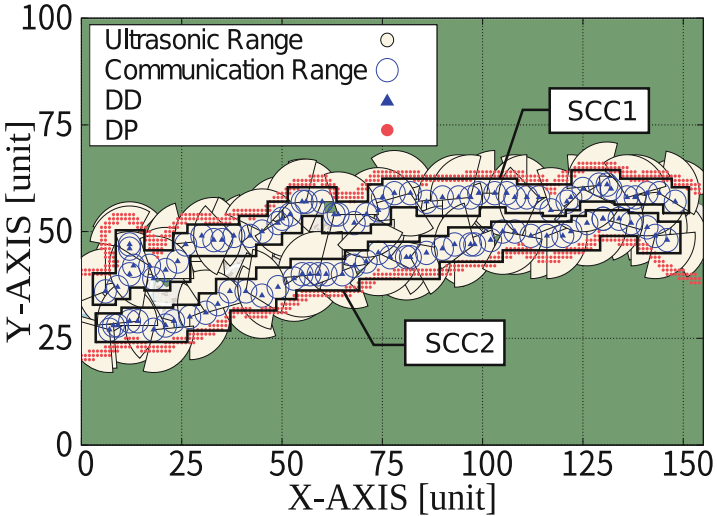


Fig. 2. The $NCDD$ and the $SSCC$ vs. the number of iterations.



(a) Initial placement.



(b) Optimized placement.

Fig. 3. Visualization results.

3 Simulation Results

The simulation settings are shown in Table 1. Figure 2 shows $NCDD$ and $SSCC$ vs. the number of iterations. In the simulation results, the number of DDs in SCC_1 is 53 and $NCDD_1$ is 440. While the number of DDs in SCC_2 is 47 and $NCDD_2$ is 382. Therefore, the $NCDD$ is 822 and 98 [%] of DPs are covered. In Fig. 3 are shown the visualization results. Figure 3(a) shows the initial placement and Fig. 3(b) shows the optimized placement.

4 Conclusions

In this paper, we proposed a SA-based *DD* placement optimization system for preventing wild deer damages. We presented the proposed simulation system and evaluated its performance by simulations. The simulation results show that the proposed system makes a good *DD* placement that maximizes *SSCC* and *NCDD*.

In the future work, we would like to improve the proposed system by minimizing the number of *DDs*.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number JP20K19793.

References

1. Laguna, E., et al.: Evaluation of a combined and portable light-ultrasound device with which to deter red deer. *Eur. J. Wildl.* **68**(4), 1–9 (2022). <https://doi.org/10.1007/s10344-022-01599-2>
2. Honda, T.: A sound deterrent prevented deer intrusions at the intersection of a river and fence. *Mammal Study* **44**(4), 267–274 (2019)
3. Fox, S., et al.: Roadkill mitigation: trialing virtual fence devices on the west coast of Tasmania. *Aust. Mammal.* **41**(2), 205–211 (2018)
4. Ranparia, D., et al.: Machine learning-based acoustic repellent system for protecting crops against wild animal attacks. In: *Proceedings of the IEEE 15th International Conference on Industrial and Information Systems (IEEE ICIIS-2020)*, pp. 534–539 (2020)
5. Blanco, J., et al.: Design and implementation of a wireless sensor and actuator network to support the intelligent control of efficient energy usage. *Sensors* **18**(6), 1892–1908 (2018)
6. Xia, F.: QoS challenges and opportunities in wireless sensor/actuator networks. *Sensors* **8**(2), 1099–1110 (2008)
7. Korber, H.: Modular wireless real-time sensor/actuator network for factory automation applications. *IEEE Trans. Industr. Inf.* **3**(3), 111–119 (2007)
8. Vassiss, D., et al.: Performance evaluation of single and multi-channel actor to actor communication for wireless sensor actor networks. *Ad Hoc Netw.* **4**(4), 487–498 (2006)
9. Hirata, A., Oda, T., Saito, N., Hirota, M., Katayama, K.: A coverage construction method based hill climbing approach for mesh router placement optimization. In: Barolli, L., Takizawa, M., Enokido, T., Chen, H.-C., Matsuo, K. (eds.) *BWCCA 2020. LNNS*, vol. 159, pp. 355–364. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61108-8_35
10. Hirata, A., et al.: A delaunay edge and CCM-based SA approach for mesh router placement optimization in WMN: a case study for evacuation area in Okayama city. In: Barolli, L., Kulla, E., Ikeda, M. (eds.) *EIDWT 2022. Lecture Notes on Data Engineering and Communications Technologies*, vol. 118, pp. 346–356. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95903-6_37
11. Oda, T., et al.: A genetic algorithm-based system for wireless mesh networks: analysis of system data considering different routing protocols and architectures. *Soft. Comput.* **20**(7), 2627–2640 (2016). <https://doi.org/10.1007/s00500-015-1663-z>

12. Oda, T., et al.: Analysis of mesh router placement in wireless mesh networks using Friedman test considering different meta-heuristics. *Int. J. Commun. Netw. Distrib. Syst.* **15**(1), 84–106 (2015)
13. Skalak, D.B.: Prototype and feature selection by sampling and random mutation hill climbing algorithms. In: *Proceedings of the 11th International Conference on Machine Learning (ICML-1994)*, pp. 293–301 (1994)
14. Kirkpatrick, S., et al.: Optimization by simulated annealing. *Science* **220**(4598), 671–680 (1983)
15. Holland, J.H.: Genetic algorithms. *Sci. Am.* **267**(1), 66–73 (1992)
16. Glover, F.: Tabu search: a tutorial. *Interfaces* **20**(4), 74–94 (1990)
17. Hirata, A., Oda, T., Saito, N., Yasunaga, T., Katayama, K., Barolli, L.: A simulation system for mesh router placement in WMNS considering coverage construction method and simulated annealing. In: Barolli, L. (ed.) *BWCCA 2021. LNNS*, vol. 346, pp. 78–87. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-90072-4_8
18. Ashtputre, S., et al.: Image segmentation based on moments and color map. In: *Proceedings of the 2nd IEEE International Conference on Communication Systems and Network Technologies (IEEE CSNT-2013)*, pp. 133–136 (2013)
19. Sa, M.: Maintenance of strongly connected component in shared-memory graph. In: *The 6th International Conference on Networked Systems*, pp. 382–387 (2018)
20. Tarjan, R.: Depth-first search and linear graph algorithms. *SIAM J. Comput.* **1**(2), 146–160 (1972)



Techno-Economic Analysis of Cloud Computing Supported by 5G: A Cloud vs on Premise Based Solutions Comparison

Christos Bouras¹(✉), Charalampos Chatzigeorgiou¹, Anastasia Kollia¹,
and Philippos Pouyioutas²

¹ Computer Engineering and Informatics, University of Patras, Patras, Greece
bouras@cti.gr, {cchatzigeorgiou, akollia}@ceid.upatras.gr

² Computer Science, University of Nicosia, Nicosia, Cyprus
pouyioutas.p@unic.ac.cy

Abstract. Cloud computing has gotten a lot of press in the IT world because it allows users to have instant access to a shared pool of configurable computer resources with no effort on their part. It is now being discussed as an enabler for more flexible, cost-effective, and powerful mobile network implementations in the communication technology (CT) sector. In this paper, a techno-economic analysis of a Cloud-based solution compared to an On-premise based one is developed. The technologies are analyzed in a technical way. Mathematical models that help determine the models' pricing is analyzed. What is more, several experiments are conducted determining if the advantages and profits outweigh the disadvantages of each proposed solution. Also, the way 5G helps scale all these processes is analyzed. Finally, this article analyzes the conclusions of the work, as well as the result of the techno-economic study that was carried out, to explain to the reader the overall benefits provided to users by exploiting the Cloud Computing technology.

Keywords: Cloud · On premise · Sensitivity analysis · 5G · Cost models

1 Introduction

Wireless networks, cloud, and mobile computing are rapidly growing in the field of Mobile Cloud Computing. With the significantly expanded limits of the fifth generation of mobile networks (5G) versatile organizations, Mobile Cloud Computing (MCC) administrations are relied upon to observe a time of the fast turn of events and become another focal point concerning portable services [1]. It is expected that individuals' work examples and ways of life in a future continually interconnected society, will be drastically altered by MCC. Future applications will be empowered by 5G, and MCC will significantly affect pretty much every part of computerized life.

The introduction of Cloud Computing is imperative in 5G, which is confirmed by research in the field. All major companies in the field of technology

(e.g. Microsoft, Google, Amazon, Yahoo, etc.) have already created their own Cloud Computing Services, which offer huge profits at a significant cost [2]. The Cloud empowers another way to send, cooperate and utilize fundamental undertaking applications. In any case, the seemingly immediate structure contributes to complexity [3].

Several papers have been released. In [4] it is pointed out that the “softwarization” of 5G is imperative and becomes a reality through new technologies, such as Software Defined Networking (SDN), Network Function Virtualization (NFV), and Cloud Computing (CC). In this context a cost model for estimating capital expenditure (CAPEX), operating expenditure (OPEX) and the total Cost of Ownership (TCO) for the proposed architecture is provided. Study [5] focuses on the comparison of many active and proposed technical pricing models and the advantages and disadvantages of each are pinpointed. The comparison is based on many aspects such as fairness, price approximation, and more. Additionally, paper [6] fosters a techno-financial structure for Cognitive Radio (CR) innovation and contrasts a current model for SDN networks.

In this paper, the main focus is the analysis and representation of the benefits of an organization/company. A cloud-based solution, for the deployment of a data center, is chosen against an on premise-based solution. The advantages of Cloud Computing are analyzed and combined with the benefits 5G networks could provide.

The remaining part of this paper is organized as follows: In Sect. 2 the proposed models/arrangements are examined and clarified. In Sect. 3 the proposed monetary models are summed up. In Sect. 4 the experimentation boundaries are selected. In Section 5 conclusions are summed up and future investigation in the field is proposed.

2 Proposed Models

The five categories (on-premise, hosted, public cloud, private cloud, and hybrid) are the main options offered alongside the installation in a corporate environment (OnPremises) or in a Cloud environment of one (hosted or Software as a Service (SaaS)). The main advantages and disadvantages of these models are identified and a more detailed comparison is proceeded, which concerns a techno economic study in a medium-sized business that is interested in choosing between On-Premise data center implementation and Cloud-based implementation through a provider.

In addition to the initial investment, there is a monthly cost for the use which concerns the license of the system (this cost is the same in both implementation scenarios) and the monthly cost of the machines that host it. It is fully explained in paper [7]. In the on-premise installation, this cost consists of the maintenance costs of the equipment and the cost of energy for its operation while in the cloud this cost is summarized in the price that the company pays to the provider of cloud services.

In the category of cloud installation, apart from the choice of a hosted solution in several cases (depending on the Cloud provider), there is the possibility of

using the system as SaaS. In this scenario, the customer is also exempted from the cost of renting the machines that will host the system and paying the provider (who is the system manufacturer himself) the price for the use of the system.

In this section, the proposed solutions are analyzed in a technical way.

2.1 Cloud Based Solution

Cloud-based solutions (or ‘cloud’ for short) stands for on demand delivery of computing resources over the Internet. On a pay-for-use-basis, you can get access to as many resources as you need such as storage space, software and applications, networks, and other on-demand services. There are three types of cloud-based systems: (IaaS): Infrastructure as a service allows you to rent storage, networks, virtual machines, servers, etc. from a cloud service provider. This is usually contracted as pay as you go. (PaaS): Platform as a service provides you with a “space” to build, deliver, test, and manage various apps. This way, you can focus on software development, instead of creating and managing the underlying infrastructure. (SaaS): Software as a service refers to the delivery of cloud-based software solutions. The cloud provider hosts the app as well as the infrastructure. They also take care of software maintenance and upgrades.

Cloud-Based solution differs from on-premises. An association has everything in-house in an on-premise arrangement, while in a cloud arrangement, an outcast provider is able to provide all that. This allows associations to provide according to circumstances and sufficient increase or decrease depending on overall usage, customer needs, and improving a correlation.

Disruptive digital technologies like cloud computing have produced a new kind of employee: the cloud worker. Cloud workers spend more than half their day working in cloud based business apps, moving seamlessly between different devices. A cloud-based worker utilizes virtual advancement to have a company’s applications offsite. There are no capital expenses, data can be upheld up regularly, and businesses simply need to pay for the resources used [8]. For those that plan strong augmentation for an overall reason, the cloud is more appealing since it grants interface to customers, associates, and various associations wherever with minimum effort. In Fig. 1, a Cloud-Based Architecture is presented.

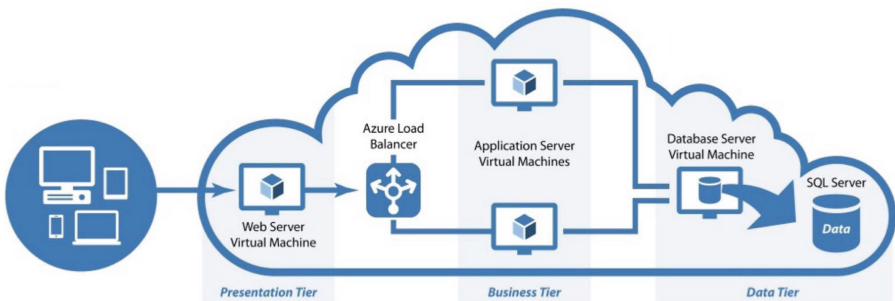


Fig. 1. Cloud based architecture

2.2 On Premise Based Solution

Whether or not a company places its applications in the cloud or regardless of whether it decides to keep them on premises, data security will be essential. However, the decision may at this point be made concerning whether to house their applications on-premise or not. Furthermore, realizing that information is situated inside in-house workers and IT framework may likewise provide more genuine feelings of serenity in any case.

On-premise programming necessitates that an endeavor buys a permit or a duplicate of the product to utilize it. Since the actual product is authorized and the whole example of programming dwells inside an association’s premises, there is by and large more prominent insurance than with a distributed computing foundation.

The disadvantage of on-premise conditions is that expenses related to overseeing and keeping up with all the arrangements involves can run dramatically higher than in a distributed computing climate. An on-premise arrangement needs in-house worker equipment, programming licenses, mixed capacities, and IT representatives close by to help and oversee potential issues that might emerge. Besides the proportion of help that an association is at risk for when something breaks or not working. In Fig. 2, an On Premise Based Architecture is introduced.

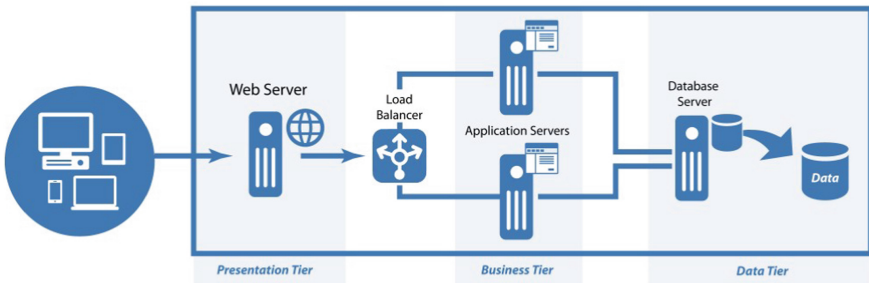


Fig. 2. On premise architecture

Figure 3 below, presents a comparison chart of on-premise versus cloud Property Management System (PMS):

	ON-PREMISE PMS	CLOUD PMS
Deployment	Data stored on a server located at the property. The program is installed on each computer from which the PMS is accessed.	Data stored on a secure, shared server at the vendor's data center. Users access the PMS through a Web browser (online).
Technical Requirements	<ul style="list-style-type: none"> • Workstations (computers) • Data server/s (ideally a dedicated server) • Compatible operating system (e.g. Microsoft Windows) • Back-up servers and hard drives • Network cards and switches/hubs • Terminal/Citrix server (for remote access and central reservations) • Interface computers 	<ul style="list-style-type: none"> • Workstations (computers or mobile devices) • Internet connection • Web browser (e.g. Google Chrome or Mozilla Firefox or Microsoft Edge)
Data Security	Responsibility of the property	Responsibility of the PMS vendor
Pricing & Costs	<ul style="list-style-type: none"> • Licencing fees (typically per workstation) • Maintenance fees • Hardware and IT infrastructure costs → CapEx	<ul style="list-style-type: none"> • A one-time setup fee • Subscription pricing (typically per room per month) → OpEx

Fig. 3. Comparison chart of on-premise versus cloud

3 Pricing Model

In this section, the proposed solutions are considered following a cost perspective. The scenario that concerns the techno-economic study is about a medium-sized enterprise that is interested in implementing an activity of Data Center and choosing between on-premise implementation and cloud based implementation through a provider. Two different solutions will be analyzed. There is a solution concerning a Cloud-Based model and another one concerning an OnPremise based model.

In every pricing model, there are a few explicit expenses. Scientifically, there are the Capital (CAPEX) and the Operational (OPEX) Expenditures [9]. The CAPEX incorporates every cost made ahead of time during the execution time in the organization. These uses incorporate a wide range of costs that are identified with the structure of the organization, for example, essential hardware, locales, and so forth. Then again, the OPEX has to do with costs that are required for the framework's day-to-day activity, the board, and coordination. The (TCO) is the aggregate sum of cash that should be paid to get a particular innovation and is the amount of CAPEX and OPEX [10].

The purpose of the analysis is to highlight the advantages and benefits of implementing a cloud-based model. So looking at the relevant research, analysis, and writing, the initial study concludes with the values in the Table below where a cost comparison is made between the two implementations. The parameters considered crucial to the calculation of CAPEX and OPEX are represented in the tables shown in Fig. 4 and 5.

$C_{server} * NoS$	The cost of each server on the total need
$C_{ne} * N_{ne}$	The cost of a each network equipment on the total need
$C_{st} * N_{st}$	The cost of each storage need on the total need
$C_{stb} * N_{stb}$	The cost of each backup storage on the total need
$C_{st(os)} * N_{st(os)}$	The cost of software on the total need
$C_{stB} * N_{stB}$	The cost of database software on the total need
$C_{st(mg)} * N_{st(mg)}$	The cost of management software on the total need
$C_{labor} * N_{labor}$	The cost of labor on the total need
$C_{estate} * N_{estate}$	The cost per square ft on the total

Fig. 4. CAPEX calculation parameters

3.1 Cloud Based Solution

1. $CAPEX = C_{server} * NoS + C_{ne} * N_{ne} + C_{st} * N_{st} + C_{stb} * N_{stb} + C_{st(os)} * N_{st(os)} + C_{stB} * N_{stB} + C_{st(mg)} * N_{st(mg)} + C_{labor} * N_{labor} + C_{estate} * N_{estate}$
2. $OPEX = C_{cp} * N_{hr} + C_{st} * N_{st} + C_{bw} * N_{bw} + C_{ss} * N_{ss} + C_{im} * N_{im} + C_{sm} * N_{sm} + C_e * N_e + C_{rent} * N_{rent} + C_{om} * N_{om} + C_{ppu} * N_{ppu}$
3. $TCO = CAPEX \text{ (Cloud Based)} + OPEX \text{ (Cloud Based)}$

3.2 On Premise Based Solution

1. $CAPEX = C_{server} * NoS + C_{ne} * N_{ne} + C_{st} * N_{st} + C_{stb} * N_{stb} + C_{st(os)} * N_{st(os)} + C_{stB} * N_{stB} + C_{st(mg)} * N_{st(mg)} + C_{labor} * N_{labor} + C_{estate} * N_{estate}$
2. $OPEX = C_{cp} * N_{hr} + C_{st} * N_{st} + C_{bw} * N_{bw} + C_{ss} * N_{ss} + C_{im} * N_{im} + C_{sm} * N_{sm} + C_e * N_e + C_{rent} * N_{rent} + C_{om} * N_{om} + C_{ppu} * N_{ppu}$
3. $TCO = CAPEX \text{ (On premises)} + OPEX \text{ (On premises)}$

$C_{cp} * N_{hr}$	The cost of computing power per kilowatt hour
$C_{st} * N_{st}$	The cost of network equipment
$C_{bw} * N_{bw}$	The storage cost per unit price
$C_{ss} * N_{ss}$	The cost of an employee's salary on the number of employees
$C_{im} * N_{im}$	The cost of infrastructure maintenance over the total cost
$C_{sm} * N_{sm}$	The cost of maintaining the software over the total cost
$C_e * N_e$	The cost of electricity per year
$C_{rent} * N_{rent}$	The cost of renting per sq.m. on the number of sq.m.
$C_{om} * N_{om}$	The cost of other maintenance per year
$C_{ppu} * N_{ppu}$	The percentage of profits from pay per use services

Fig. 5. OPEX calculation parameters

4 Parameter Selection

The table in Fig. 6 incorporates every boundary and factor that is identified with the valuing models. These variables are based on previous research activities and are thoroughly explained in the tables shown in Fig. 4 and Fig. 5. Also, value ranges are decided on the SA, that is utilized for the exploratory examination. SA is a notable method, where a few boundaries of an item are broken down and it is shown if they influence a financial model and how much effect they have on this model. This procedure helps demonstrate which network boundaries ought to be diminished.

4.1 Cost Comparison

In this section, the cost comparison of the two different models is presented similarly to the way shown in [11] and also in [12]. Studying the Table in Fig. 6 it is noted that different costs for both implementations can significantly affect the final cost [13], in both CAPEX and OPEX since, if they are looked separately, their importance is better understood [14].

Examples of these cost-dependent factors are displayed in the two sections following.

4.2 On Premises Solution Cost Adjustments

As seen in Fig. 7, the cost of Servers increases proportionally if there emerges a need for more servers. Which in turn causes CAPEX costs to increase accord-

	On-premises				Cloud Based			
	Unit	Quantity	Per Unit	Total	Unit	Quantity	Per Unit	Total
CAPEX (initial)				3.110.000,00 €				90.000,00 €
Server	Nos	100	4.000 €	400.000 €	Nos	0	0 €	0 €
Network Equipments	Nos	50	1.000 €	50.000 €	Nos	0	0 €	0 €
Storage	TB	50	3.500 €	175.000 €	TB	0	0 €	0 €
Storage (Backup)	TB	350	1.500 €	525.000 €	TB	0	0 €	0 €
Software (OS + IIS)	Nos	100	2.500 €	250.000 €	Nos	0	0 €	0 €
Software (DB)	Nos	100	15.000 €	1.500.000 €	Nos	0	0 €	0 €
Software (AV + Mgmt)	Nos	100	300 €	30.000 €	Nos	0	0 €	0 €
Labor for Start	€/resource	10	18.000 €	180.000 €	€/resource	5	18.000 €	90.000 €
Real Estate	€/sft	0	1.000 €	0 €	€/sft	0	0 €	0 €
OPEX (annual)				999.000 €				717.000,00 €
Computing Power	€/hr	0	0 €	0 €	€/hr	3000000	0,16 €	480.000,00 €
Storage	TB	0	0 €	0 €	€/GB	0	0,12 €	0,00 €
Bandwidth	€/annum	3	20.000 €	60.000 €	€/GB	0	0,18 €	0,00 €
Staff Salary	staff/annum	8	28.000 €	224.000 €	staff/annum	4	28.000 €	112.000,00 €
Infrastructure Maintenance	% of total cost	35	7.000 €	245.000 €	% of total cost	0	0 €	0,00 €
Software Maintenance	% of total cost	35	5.000 €	175.000 €	% of total cost	0	0 €	0,00 €
Electricity	€/annum	1	90.000 €	90.000 €	€/annum	0	0 €	0,00 €
Rent for Real Estate	€/sft/annum	1000	100 €	100.000 €	€/sft/annum	0	0 €	0,00 €
Other Maintenance	€/annum	3	35.000 €	105.000 €	€/annum	0	0 €	0,00 €
Pay-per-Use Savings	%	0	0 €	0 €	%	25	5.000 €	125.000,00 €
TOTAL				4.109.000,00 €				807.000,00 €
Savings				3.302.000,00 €				

Fig. 6. Cost comparison

ingly as seen in Fig. 8. Thus considered, the plan on the server needs should be extensive and forward-looking in order to avoid future costs depended on server units.

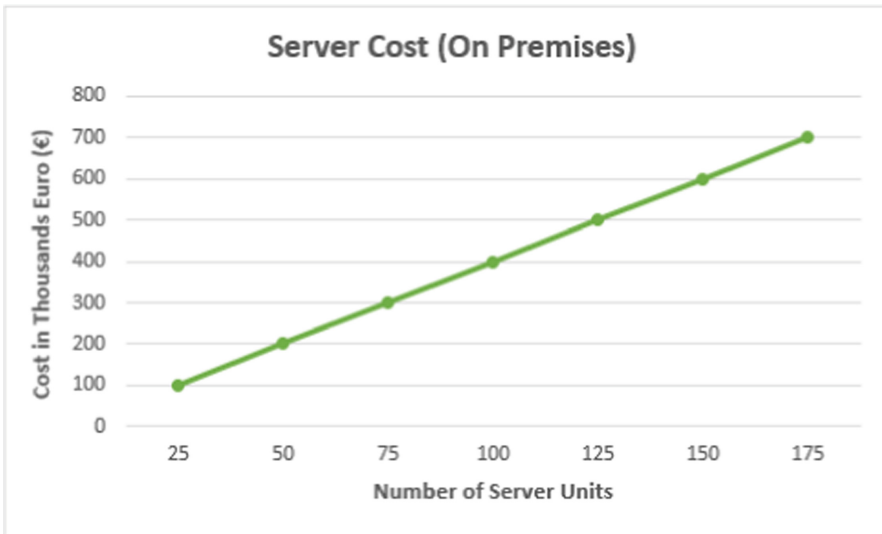


Fig. 7. Server cost impact

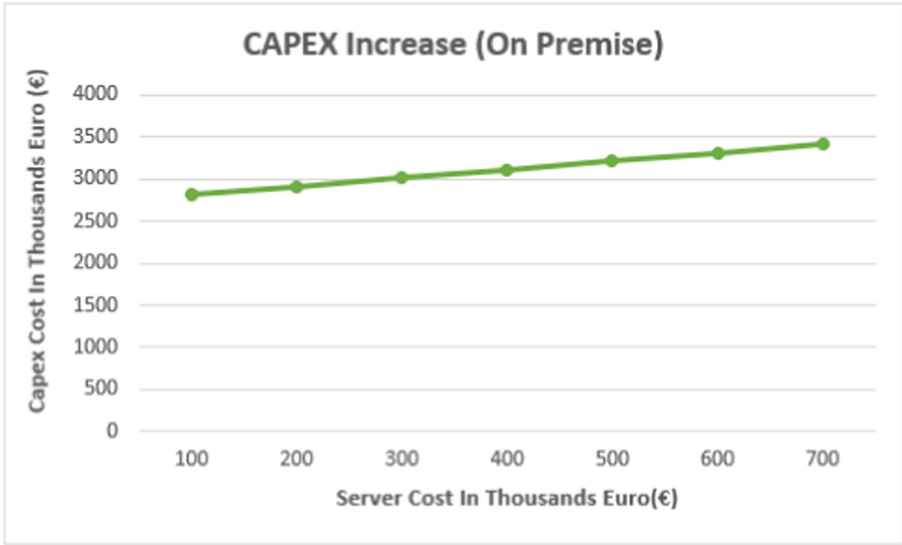


Fig. 8. CAPEX cost increase due to server cost

4.3 Cloud Based Solution Cost Adjustments

As seen in Fig. 9, the cost of Computing Power is directly related to the price per kilowatt-hour (kWh) which may differ in each operating range. Which in turn can lead to significant increases in the overall OPEX, as shown in Fig. 10.

4.4 Performance Evaluation

Based on the above described facts, a very detailed research and market investigation should take place before the final decision because the cost per kilowatt-hour varies between places and even cities is over 4 times larger than in the Cloud-Based solution we propose. The Cloud-based solution offers more adaptability and versatility and is less dependent on unpredictable factors. The results are displayed in Fig. 11 and Fig. 12 for consideration. It is finally found that the difference in the two implementations is significant in terms of cost, especially if this is observed in the depth of 4 years. Cloud-Based implementation is significantly superior to the On-Premise, something that is reflected in Fig. 9. To summarize the experimental comparison, it is obvious that a Cloud-based solution is more efficient in the final choice. The cost of an On-Premise solution varies and is strongly dependent on multiple factors that are not always controllable by the side that implements the solution. Surely the fact the cost of both solutions may be less after 4 or 5 years is something to be considered in the long run, but in a shorter-term case, the Cloud-based solutions dominate over the On-Premise based ones.

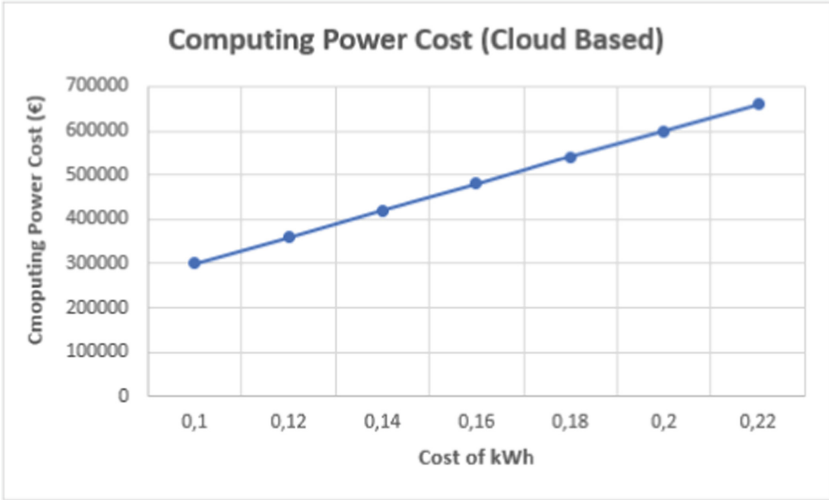


Fig. 9. Computing power cost impact

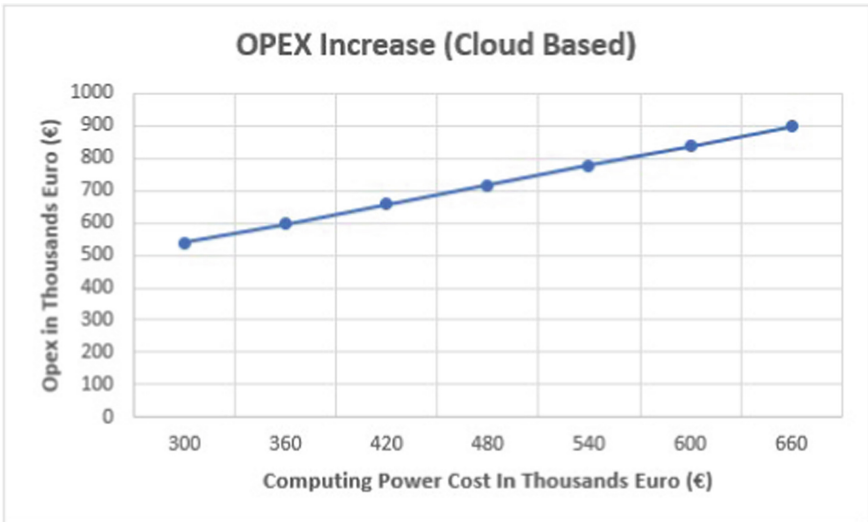


Fig. 10. OPEX cost increase due to computing power cost

	On-premises			Cloud Based		
	CAPEX	OPEX	Cost	CAPEX	OPEX	Cost
YEAR - 1	3.030.000 €	999.000 €	4.029.000 €	90.000 €	717.000 €	807.000 €
YEAR - 2	-	999.000 €	5.028.000 €	-	717.000 €	1.524.000 €
YEAR - 3	-	999.000 €	6.027.000 €	-	717.000 €	2.241.000 €
YEAR - 4	-	999.000 €	7.026.000 €	-	717.000 €	2.958.000 €

Fig. 11. 4-year plan cost estimates

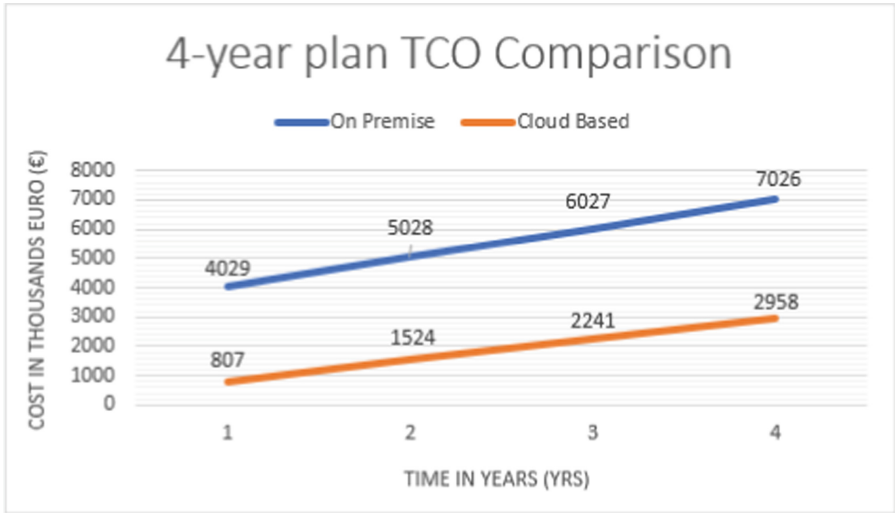


Fig. 12. 4-year plan TCO comparison

5 5G and Cloud Computing

What 5G brings to the above though? Up to here, we evaluated the benefits cloud computing offers against On-premise solutions. But how 5G is related to that and why is mentioned in this paper? The combination of 5G speed with cloud computing’s powerful tools and flexibility is likely to herald a new generation of computing capability. Shift or moving towards cloud storage and cloud computing is a developing and fundamental peculiarity nowadays and in these pandemic conditions, its need is more crucial. Information Technology (IT) is at a critical juncture in history, when the transformation to a digital future is unavoidable, in the technological era in which we now live. As of today, more than half of the world’s population is linked, which is amazing, but it is only the tip of the iceberg. Through 5G and beyond networks, “everything that can be connected will be connected”.

Regarding Cloud Computing now, it simply means bringing cloud technology closer to the end-user to reduce latency, boost downstream bandwidth, and decrease upstream bandwidth. This allows telecommunication providers to deliver services considerably more quickly, giving them the agility they need to compete. Cloud computing allows for faster analysis and response times by utilizing automation tools that act on local data. With the introduction of 5G, all network/telco services will operate similarly to cloud-based services, benefiting from deployment agility, scalability, and other benefits [15].

In terms of some actual benefits of the 5G introduction to Cloud Computing, wireless connectivity environments, such as factories, have largely relied on wired technologies to get the appropriate network reliability, predictability, and latency characteristics [16]. Wired networking technologies are expensive to install, require real estate, and need maintenance. Private 5G has the potential to replace wired technologies in these sensitive environments. Moreover, environments such as farms, oil fields, and mines may not even have connectivity, to begin with since they are not well suited to wired technologies. Such environments can take advantage of Private 5G for solid networking connectivity. The next compelling benefit of 5G+ cloud computing is the promise to save operating expenses (OPEX). For a factory, this could be via robotics control, autonomous vehicles, AI/ML (Artificial Intelligence/Machine Learning) quality inspection, IoT management, and more. For hospitals, it could be through radiology anomaly detection at the edge. For precision agriculture, it could be via drone control and IoT management. Video surveillance applications could be used for retail store security to slash costs. Smart building applications could cut energy costs and optimize space utilization with 5G+.

6 Conclusions and Future Work

In this paper, a comparative study was presented between the two different implementations for the creation and operation of a data center. The advantages of a CloudBased architecture in combination with the capabilities of 5G networks were presented, which incorporate various state-of-the-art technologies. With the proposed architecture the disadvantages of on-premise implementation are mainly summarized in the financial costs of this option.

The comparative study leads to the conclusion that cloud based implementation is more economical than on-premise since, for this example, it generates a profit of €4,000,000. Even if a significantly larger amount is added to the cloud based implementation - of the order of €500,000 indicatively - as migration cost, again the difference in costs, especially in a four-year plan as described above, is significant.

In future research aspects of increased TCO in CloudBased solutions must be researched according to opinions saying that after 5 or 6 years on a Cloud-Based solution the cost may increase a lot, meeting On-Premises solutions finally. This comes as a result of the increasing yearly cost of Cloud Storage needed for the purposes [17].

Cloud checking as a Service requires more investigation and evaluation to gain a better understanding of cloud facilitated apps. There are opportunities for improvement in the areas of up time, consistency, weaknesses, occurrences, combination, and so on. Managing the use of multi-cloud framework administrations will become increasingly important in the future. Especially if On-Premise is doomed in the long term due to suppliers' narrow-minded motivations.

Cloud-based arrangements, namely SaaS arrangements, provide a flexible and accessible option for accessing continual data whenever and wherever it is convenient. Because they rely on a pay-as-charges-arise evaluating strategy, the association/organization benefits from lower upfront costs for equipment and programming, as well as adaptability that allows us to change the arrangement as the company demands change.

References

1. Huth, A., Cebula, J.: The basics of cloud computing (2011)
2. Shi, Z.: Cloud computer architecture based on enterprise-class cloud model and its key technologies research. *Int. J. Res. Bus. Stud. Manag.* 5–11 (2016)
3. Drobik, A., Maoz, M.: Adapting your it strategy for a cloud-dominated business application environment. *Gartner Research Gate* (2016)
4. Bouras, C., Ntarzanos, P., Papazois, A.: Cost modeling for SDN/NFV based mobile 5G networks. In: 2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 56–61 (2016)
5. Al-Roomi, S.B., Al-Ebrahim, S., Ahmad, I.: Cloud computing pricing models: a survey. *Int. J. Sci. Eng. Res.* 22–26 (2019)
6. Bouras, C., Kollia, A., Papazois, A.: Techno-economic analysis of cognitive radio models in 5G networks. *Procedia Comput. Sci.* **175**, 300–307 (2020)
7. Chandra, D.G., Borah, M.D.: Cost benefit analysis of cloud computing in education. In: 2012 International Conference on Computing, Communication and Applications, pp. 1–6 (2012)
8. Hernandez, J., et al.: Comprehensive model for technoeconomic studies of next-generation central offices for metro networks. *J. Opt. Commun. Netw.* **12**(12), 414–427 (2020)
9. Liew, S.H., Su, Y.Y.: CloudGuide: helping users estimate cloud deployment cost and performance for legacy web applications. In: 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, pp. 90–98 (2012)
10. Aggarwal, S., McCabe, L.: A 4-year total cost of ownership (TCO) perspective comparing cloud and on-premise business application deployment. *The Compelling TCO Case for Cloud Computing in SMB and Mid-Market Enterprises*. Hurwitz & Associates White Paper (2009)
11. Jin, Y., Wen, Y., Guan, K., Kilper, D., Xie, H.: Toward monetary cost effective content placement in cloud centric media network. In: 2013 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2013)
12. Gedel, I., Nwulu, N.: Infrastructure sharing for 5G deployment: a techno-economic analysis. *Int. J. Interact. Mob. Technol. (iJIM)* **15** (2021)
13. Juhasz, Z.: Quantitative cost comparison of on-premise and cloud infrastructure based EEG data processing. *Cluster Comput.* **24**, 625–641 (2021)
14. Fisher, C.: Cloud versus on-premise computing. *Am. J. Ind. Bus. Manag.* **08** (1991)

15. Hernandez, J.A., Quagliotti, M., Serra, L.: On the cloudification of metropolitan area networks: impact on cost and energy consumption. In: 2021 IEEE 7th International Conference on Network Softwarization (NetSoft), pp. 330–338 (2021)
16. Paglierani, P., et al.: Techno-economic analysis of 5G immersive media services in cloud-enabled small cell networks: the neutral host business model: providing techno-economic guidelines for the successful provision of 5G innovative services in small cell networks. *Trans. Emerg. Telecommun. Technol.* **31** (2020)
17. Arokia, S.S.R., Rajan, P.: Evolution of cloud storage as cloud computing infrastructure service. *J. Comput. Eng. (IOSRJCE)* 38–45 (2012)



An Integrated Fog-VDTN Architecture for Data Dissemination

Evjola Spaho^(✉)

Department of Electronics and Telecommunication, Faculty of Information Technology, Polytechnic University of Tirana, Mother Teresa Square, No. 4, Tirana, Albania
espaho@fti.edu.al

Abstract. In this paper, an integrated architecture based on fog computing and Vehicular Delay Tolerant Networks (VDTNs) for data dissemination is proposed. Fog nodes update their content from the cloud and use VDTNs to deliver their content to other nodes. VDTNs enable communication in low density networks where connectivity is low by using store-carry-forward approach. This approach is suitable for non-urgent data. The vehicles mobility will be used to send data from the fog nodes to all the other nodes in the network. Fog nodes realize storing and computing functions and are stationary nodes scattered in the vicinity where the vehicles move. Fog nodes will disseminate different data like advertisements or flyers with information to other nodes. Usage of this integrated architecture of fog nodes and VDTNs reduces the cost for data transmission and saves communication resources.

1 Introduction

Recently, the amount of data in vehicular networks is increased significantly because of the exponential expansion of the generated vehicular data, the increase of vehicle numbers, and the rise of in-car user data demands.

Various types of smart cameras and sensors, wireless communication modules, storage, and compute resources can now be installed in vehicles thanks to advancements in automotive technology. Massive volumes of data are produced from tracking the on-road and on-board condition. Also, many smart devices, smart cameras and sensors are integrated into vehicles.

Vehicles are able to connect and exchange crucial information in vehicular networks. Vehicular networks can be used for different applications for road safety, traffic management or infotainment.

Advertisement dissemination and infotainment applications produce non critical time data and sending the information with a low cost can be done using Vehicular Delay Tolerant Networks (VDTNs). VDTNs are a unique subset of Delay Tolerant Networks (DTNs) [1] where the nodes are vehicles [2–6]. VDTNs are characterized by the absence of a continuous path between source and destination and data is sent via a message switching mechanism. The performance

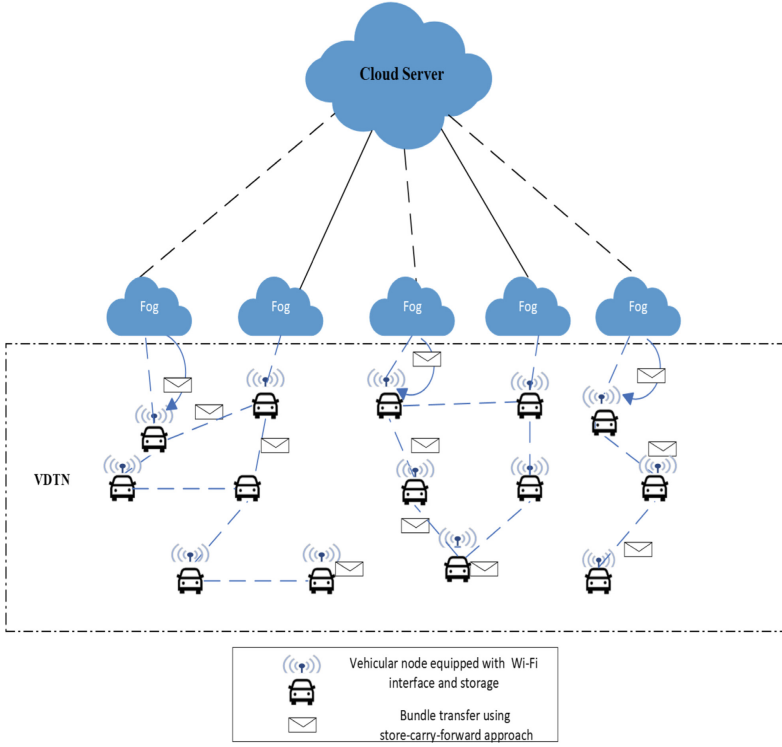


Fig. 1. The proposed Fog-VDTN architecture.

of VDTNs is strongly connected with nodes mobility. Mobility enables opportunistic contacts between nodes and data transmission.

Cloud computing offers services such as Infrastructure As A Service (IAAS), Platform As A Service (PAAS), and Software As A Service (SAAS). Cloud computing uses remote servers instead of local servers to perform storage, management and data processing.

Fog computing [7] is a decentralized computing architecture and is an extension of cloud computing that brings the cloud closer to the end-devices. Fog computing is suitable for mobile users and location-aware services and applications.

In this work is proposed the integration of VDTNs with fog computing for non urgent data dissemination. Advertisement dissemination is a delay tolerant application. For big amount of data and time non-critical data, VDTNs and message switching approach is a good solution because the cost for data transmission is reduced and communication resources are saved.

The remainder of this paper is as follows. The proposed architecture is presented in Sect. 2. The simulation system and scenarios are described in Sect. 3. In Sect. 4 are shown the simulation results. Finally, the conclusions are presented in Sect. 5.

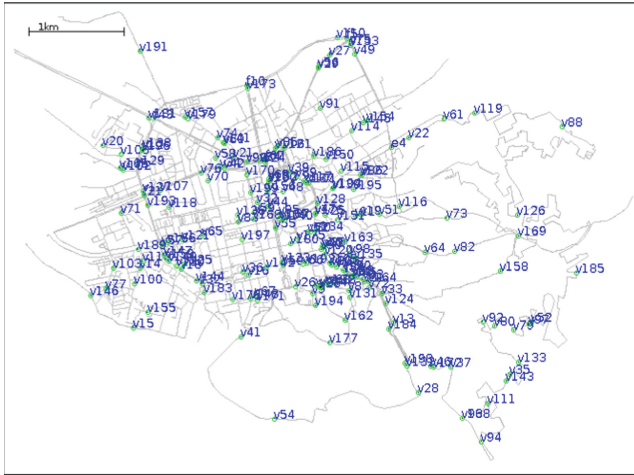


Fig. 2. Snapshot of the simulation.

2 Proposed Architecture

In the proposed Fog-VDTN architecture, fog nodes are connected with the cloud to update their content and VDTN is used for data dissemination using message switching approach.

Fog computing can be used to provide location-aware services and the data can be accessed off-line as some amount of data are stored in a local data center. Fog computing is recommended for large scale networks where fog nodes at different locations are connected with the same cloud to cover a wide region. Fog nodes are always connected with the Internet, they send and receive data from the Cloud by using wired interface.

Fog nodes will offer localized content dissemination and will distribute different advertisements to vehicles passing nearby and within their transmission range. The fog nodes are placed in strategic points of interests where most of the vehicles tend to go and they send data to all other vehicles using Wi-Fi Interfaces. Vehicles equipped with on board units that move according to map based mobility model will send this information to other vehicles by using store-carry-forward mechanism. The proposed approach uses the memory and interfaces of smart devices like on board units to form a VDTN and exchange data by opportunistic contacts as shown in Fig. 1.

3 Simulation System and Scenarios

To evaluate the performance of the proposed architecture, a JAVA based open source simulator called the Opportunistic Network Environment (ONE) [8] is used. This simulator offers the possibility to use different mobility models and

Table 1. Simulation parameters and their values.

Parameters	Values
Number of Stationary Fog Nodes	8 nodes
Number of Total Nodes	200, 300 nodes
Simulation Time	28800 s
Map Size	4 km \times 5 km
Movement Model	Map-based
Buffer Size for Fog Nodes	1000 MB
Buffer Size for Other Nodes	100 MB
Interface Type	Simple Broadcast Interface
Interface Transmission Speed	250 Kbps
Interface Transmission Range	20 m
Message TTL	30, 60, 100, 200, 300, 400 min
Vehicles Speed	5–25 km/h
Message Size	10k, 50k
Message Creation Interval	30 s

import real maps for realistic scenarios. For the simulations is considered an urban area of Tirana city in Albania. In Fig. 2 is presented a snapshot of the network at the beginning of the simulation with The ONE simulator.

Two different scenarios are implemented. In the first scenario, data from 8 fog nodes are sent to 192 vehicles moving in the streets of Tirana (for example different flyers for shopping mall sales). In the second scenario, the number of fog nodes is considered 8 and the number of vehicles is considered 292. Fog nodes are equipped with buffers of 1000 MB, transmission range 20 m. In Table 1 are listed all simulation parameters and their values.

The routing protocols used for evaluation are presented in following.

- **Epidemic Routing Protocol**

In Epidemic [9] protocol, each message is spread in the network with no priority and no limit using flooding mechanism. When two nodes encounter each other, they exchange and compare the list of their messages to find the messages that are not already in the storage of the other nodes.

- **Spray and Wait Routing Protocol**

Spray and Wait [10] uses the spray phase and the wait phase. When a new message is created in the network, a maximum of L number of copies of the message is created in the network. In the spray phase, the source of the message will spray one copy of this message to L different “relays”. When a relay receives the copy, it enters the wait phase, where it will hold that message until encounter the destination directly.

For evaluation, we use the following metrics.

- **Delivery Success Rate**

It is calculated as the ratio of number of delivered messages over the created ones.

- **Overhead Ratio**

It is calculated as the difference between relayed and delivered messages over the number of delivered messages.

- **Average Latency**

It is calculated as the average time elapsed from the creation of the messages at source to their successful delivery to the destination.

4 Simulation Results

In Fig. 3 are shown the simulation results of considered protocols for delivery success rate vs. TTL for scenarios with 200 and 300 nodes. From the simulation results, Epidemic routing protocol performs better than Spray and Wait in terms of delivery success rate. By increasing TTL from 30 min to 200 min, the performance of both protocols is improved. For TTL values longer than 200 min, for both scenarios, there is not further improvement of the performance for both routing protocols.

The results for overhead ratio are presented in Fig. 4. Epidemic has higher overhead ratio compared with Spray and Wait. The increase of TTL does not effect the overhead ratio of Epidemic. For Spray and Wait, the increase in TTL decreases the overhead ratio. The number of nodes effects the performance of Epidemic protocol in terms of overhead ratio. The bigger the number of nodes in the network is, the higher is the overhead ratio. The increase of number of nodes in the network does not effect the overhead ratio for Spray and Wait protocol.

The average delay results for both scenarios are presented in Fig. 5. The increase of TTL increases also the average delay for both protocols. The average delay is higher for the scenario with 200 nodes for both protocols. In both scenarios, Epidemic protocol performs better than Spray and Wait.

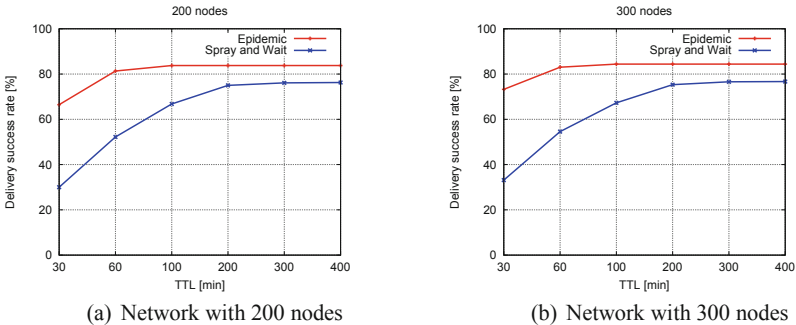


Fig. 3. Results for Delivery success rate vs. TTL.

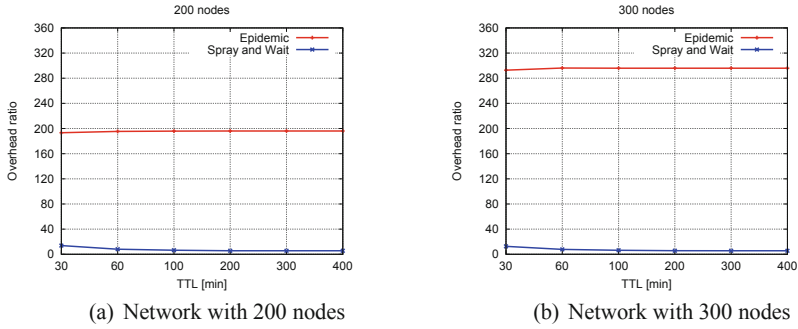


Fig. 4. Results for Overhead ratio vs. TTL.

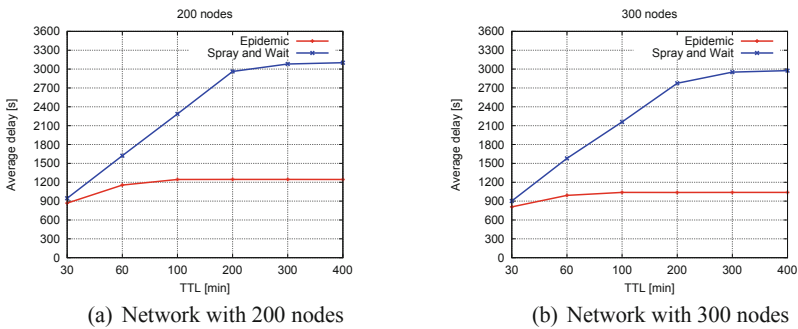


Fig. 5. Results for Average delay vs. TTL.

5 Conclusions

In this paper, an integrated architecture based on fog and VDTNs for data dissemination is proposed and evaluated by simulations. Usage of fog nodes and VDTNs reduces the cost for data transmission and saves communication resources. The performance was evaluated for Epidemic and Spray and Wait routing protocols.

In terms of delivery success rate:

- Epidemic protocol performs better than Spray and Wait.
- The increase of TTL to 200 min improves the delivery success rate for both protocols.
- The increase of number of nodes in the network has a very small effect on the delivery success rate for both protocols.

In terms of overhead ratio:

- Spray and Wait has lower overhead ratio compared to Epidemic.
- The increase of TTL does not effect the performance of both routing protocols.

- The increase of number of nodes increases also the overhead ratio for Epidemic, but it remains almost the same for Spray and Wait.

In terms of average delay:

- Spray and Wait has higher delay compared with Epidemic.
- The increase of TTL, increases also the average delay for Spray and Wait.
- The increase of number of nodes slightly increases the average delay for both protocols.

In the future, we would like to consider other routing protocols and new data collection scenarios and evaluate their performance.

References

1. Spaho, E.: Comparing different DTN routing protocols in a dense deployment scenario with realistic mobility trace. *Int. J. Innov. Technol. Interdisc. Sci.* **2**(4), 290–296 (2019). <https://doi.org/10.15157/IJITIS.2019.2.4.290-296>
2. Spaho, E., Korovesi, A.: A low-cost solution for smart-city based on public bus transportation system using opportunistic IoT. In: Barolli, L., Kulla, E., Ikeda, M. (eds.) *EIDWT 2022. LNDECT*, vol. 118, pp. 175–182. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95903-6_19
3. Spaho, E.: Usage of DTNs for low-cost IoT application in smart cities: performance evaluation of spray and wait routing protocol and its enhanced versions. *Int. J. Grid Util. Comput.* **12**(2), 173–177 (2021)
4. Bylykbashi, K., Spaho, E., Barolli, L., Xhafa, F.: Routing in a many-to-one communication scenario in a realistic VDTN. *J. High Speed Netw.* **24**(2), 107–118 (2018)
5. Spaho, E., Dhoska, K., Bylykbashi, K., Barolli, L., Kolici, V., Takizawa, M.: Performance evaluation of energy consumption for different DTN routing protocols. In: Barolli, L., Kryvinska, N., Enokido, T., Takizawa, M. (eds.) *NBiS 2018. LNDECT*, vol. 22, pp. 122–131. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-98530-5_11
6. Spaho, E., Dhoska, K., Barolli, L., Kolici, V., Takizawa, M.: Enhancement of binary spray and wait routing protocol for improving delivery probability and latency in a delay tolerant network. In: Barolli, L., Hellinckx, P., Enokido, T. (eds.) *BWCCA 2019. LNNS*, vol. 97, pp. 105–113. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-33506-9_10
7. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the Internet of Things. In: *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, Helsinki, Finland, 13–17 August 2012*, pp. 13–16 (2012)
8. Keranen, A., Ott, J., Karkkainen, T.: The ONE simulator for DTN protocol evaluation. In: *Proceedings of the 2nd International Conference on Simulation Tools and Techniques (SIMUTools 2009)* (2009). <http://www.netlab.tkk.fi/tutkimus/dtn/theone/pub/theonesimutools.pdf>
9. Vahdat, A., Becker, D.: Epidemic routing for partially connected ad hoc networks. Technical report CS-200006, Duke University (2000)
10. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In: *Proceedings of ACM SIGCOMM 2005 - Workshop on Delay Tolerant Networking and Related Networks (WDTN 2005)*, Philadelphia, PA, USA, pp. 252–259 (2005)



Energy-Consumption Evaluation of the Tree-Based Fog Computing (TBFC) Model

Dilawaer Duolikun¹(✉), Shigenari Nakamura³, Tomoya Enokido²,
and Makoto Takizawa¹

¹ Research Center for Computing and Multimedia Studies (RCCMS),
Hosei University, Tokyo, Japan

dilewerdolkun@gmail.com, makoto.takizawa@computer.org

² Faculty of Business Administration, Ritssho University, Tokyo, Japan
eno@ris.ac.jp

³ Tokyo Metropolitan Industrial Technology Research Institute, Tokyo, Japan
nakamura.shigenari@iri-tokyo.jp

Abstract. It is critical to reduce the energy consumption of information systems to realize green societies. The IoT (Internet of Things) is so scalable that millions to billions of computers and devices are interconnected in types of networks and accordingly huge amount of energy is consumed. In our previous studies, the TBFC (Tree-Based Fog Computing) model is proposed to energy-efficiently realize the IoT, where fog nodes are hierarchically structured and application processes are distributed to not only servers in clouds but also fog nodes. The energy consumption of fog nodes in the TBFC model is obtained for a collection of sensor data simultaneously issued by device nodes in the evaluation. In this paper, we evaluate the TBFC model in terms of total energy consumption of nodes where each device node periodically sends sensor data. In the evaluation, we show the energy consumption of the TBFC model is smaller than the cloud computing (CC) model of the IoT.

Keywords: Green computing systems · TBFC (Tree-Based Fog Computing) model · IoT · Fog computing (FC) model · Energy consumption

1 Introduction

The IoT (Internet of Things) is now one of the most important infrastructure to realize various applications in our societies. The IoT is so scalable that millions to billions devices are interconnected in addition to servers and clients and accordingly huge amount of energy is consumed [1, 2]. In order to decrease carbon footprint on the earth, the total electric energy consumption of the IoT has to be reduced. There are models to realize the IoT; cloud computing (CC) [3] and fog

computing (FC) [4,5] models. In the CC model [3], device nodes supporting sensors and actuators are interconnected with clouds of servers in networks. Device nodes send sensor data to servers in the clouds through networks and the sensor data is processed by the servers. Servers and networks are heavily loaded to process and transmit sensor data from huge number of device nodes. The FC model [4,5] is now widely used to efficiently realize the IoT. Here, application processes and databases are distributed to not only servers in clouds but also fog nodes. Sensor data is processed by fog nodes and processed data is mostly smaller than the sensor data like some data selected and an aggregate value like average one. Hence, the traffic of servers and networks can be reduced. On the other hand, fog nodes consume energy to process sensor data in addition to servers in the FC model while only servers consume energy to process sensor data in the CC model. In order to reduce the energy consumption of the FC model, the TBFC (Tree-Based Fog Computing) model [27,29–32] is proposed, where fog nodes are structured in a tree and sensor data is distributed to multiple fog nodes and in parallel processed by the fog nodes.

The *macro-level* power consumption and computation models [7–10,12,13,15–17,19,20,23] are proposed, which give how much electric power [W] to be consumed by a whole computer to perform application processes. By taking advantage of the models, the live migration approach of virtual machines [7–10,14,19,21,22,24] is discussed to reduce the energy consumption of servers in a cloud. Energy-aware algorithms [11,15–17,25,26] are also proposed to select a host server to perform an application process issued by a client. Power consumption and computation models of fog nodes in the IoT [29–33] are proposed. Here, the power consumption of each fog node to receive and process sensor data and to send the processed data can be obtained.

Compared with the CC model where every sensor data is received and processed by servers, the total energy consumption of the IoT can be reduced in the TBFC model. The TBFC model is evaluated in papers [29–33] where a collection of device nodes once send sensor data to edge nodes. After sending the output data to a parent node, the node gets idle while the ancestor nodes are active to calculate output data on input data. Even an idle node consumes power [8]. A node receives succeeding input data from child nodes and device nodes after sending the output data to the parent node. In this paper, we evaluate the TBFC model where device nodes periodically send sensor data in terms of the total energy consumption of servers and nodes and the delivery time of each sensor data compared with the CC model. We show the total energy consumption and delivery time of the IoT can be reduced in the TBFC model compared with the CC model in the simulation.

In Sect. 2, we present the TBFC model. In Sect. 3, the power consumption and computation models of a fog node are discussed. In Sect. 4, we evaluate the TBFC model.

2 The TBFC Model

In the CC (Cloud Computing) model [3] of the IoT, device nodes supporting sensors send sensor data to servers in clouds through networks. Application processes on the servers receive sensor data and obtains actions from the sensor data to be performed on actuators of devices. The sensor data and processed data like actions are stored in databases on the servers. The servers send the actions to device nodes and then actuators in the device nodes are activated by the actions. The FC (Fog Computing) model [4,5] of the IoT is composed of fog nodes in addition to clouds of servers and device nodes. Application processes and databases are distributed to not only servers but also fog nodes. Sensor data is processed by fog nodes and processed sensor data is sent to servers. Thus, the traffic of the servers and networks and time to activate actuators can be reduced in the FC model.

The TBFC model [27–32] is proposed to energy-efficiently realize the IoT. The TBFC model is composed of nodes which are tree-structured. A *root* node stands for a cloud of servers. A leaf node is an *edge* node which communicates with device nodes. Non-root nodes are *fog* nodes [4,5] which support not only routing functions but also application processes to handle sensor data.

A root node f is interconnected with *child* fog nodes f_1, \dots, f_b ($b \geq 0$). Each node f_i is also interconnected with child nodes f_{i1}, \dots, f_{i,b_i} ($b_i \geq 0$). Let I be a sequence $\langle i_1 i_2 \dots i_{l-1} \rangle$ ($l \geq 1$) of indexes. That is, $I (= \langle i_1 \dots i_{l-1} \rangle)$ means a sequence $\langle f, f_{i_1}, f_{i_1 i_2}, \dots, f_{i_1 i_2 \dots i_{l-1}} \rangle$ of nodes which is a path from the root node f to the parent node $f_{i_1 i_2 \dots i_{l-1}} (= f_I)$ of the node f_{Ii} . $|I|$ shows the number $l - 1$ of indexes in I . A node f_I is at level $|I| + 1$. For example, a root node f and a fog node f_{21} are at levels 1 and 3, respectively. Thus, each node f_{Ii} communicates with a parent node f_I and child nodes $f_{Ii1}, \dots, f_{Ii b_i}$ ($b_i \geq 0$). A leaf node f_I is named *edge* node which communicates with a *device* node s_I which is equipped with sensors or actuators. Here, the device node s_I sends sensor data to the edge node f_I and the edge node f_I sends actions to the device node s_I .

Each node f_I supports a process p_I which is an $m_I (> 0)$ -ary function $p_I(x_1, \dots, x_{m_I})$ where each parameter x_i is typed D_{Ii} and takes a collection of data of type D_{Ii} . The output data x of the function $p_I(x_1, \dots, x_{m_I})$ is typed D_I . The node f_I supports m_I input ports $ip_{I1}, \dots, ip_{Im_I}$ and one output port op_I as shown in Fig. 1. A node f_I receives input data at an input port ip_{Ii} from a child node f_{Ii} and passes the input data to the process p_I as the i th parameter x_i . Here, a type of an input port ip_{Ii} is a type D_{Ii} of the i th parameter x_i of the process p_I . A node f_I receives input data id_{Ii} at the input port ip_{Ii} , which is sent through the output port op_{Ik} of a child node f_{Ik} . Here, the type of the output port op_{Ik} has to be the same as the type of the input port ip_{Ii} . Thus, a pair of the parent node f_I and child node f_{Ik} are connected in an output-input ($op_{Ik} \rightarrow ip_{Ii}$) relation. For each input port ip_{Ii} of a node f_I , multiple child nodes can be connected. Let CF_{Ii} be a set of child nodes of a node f_I , whose output ports are connected to an input port ip_{Ii} of the node f_I . Let ID_{Ii} be a set of input data sent to an input port ip_{Ii} from child nodes in the set CF_{Ii} . There are types of input ports. If an input port ip_{Ii} is a *conjunctive* type, a fog node

f_I blocks until every child node in the set CF_{Ii} sends input data. Then, all the input data in ID_{Ii} are passed to a process p_I as the i th parameter. On the other hand, if an input port ip_{Ii} is a *disjunctive* type, once input data is received from a child node, the input data is passed to the process p_I and input data from the other child nodes are neglected. If a node f_I receives input data at every input port, the process p_I calculates the output data od_I on the input data. The node f_I sends the output data od_I through the output port op_I to the input port of the parent node. Thus, the root node f finally receives the input data id_i , i.e. output data od_i from each child node f_i and obtains the output data od by calculating on the input data ID . The output data od is stored in the database and is delivered to device nodes in the tree. Here, the output data od shows an action to be performed on the device.

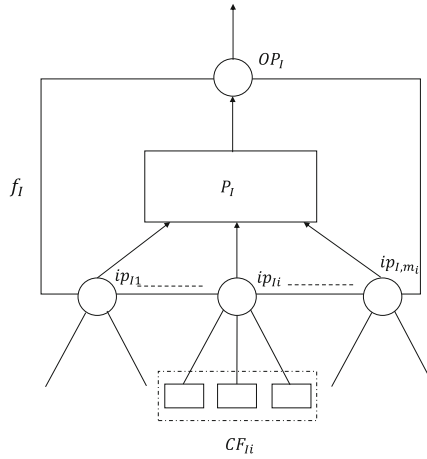


Fig. 1. Node of the TBFC model.

[Example]. Figure 2 shows an example of the TBFC model [33]. Here, there are two types of device nodes; *t-sensor* nodes s_1 and s_2 and *h-sensor* nodes s_3 and s_4 . The *t-sensor* nodes s_1 and s_2 collect temperature data every one second and send the collected data to parent edge nodes f_{111} and f_{112} , respectively, each of which supports a process *t-aggregate*. Each *t-aggregate* node f_{11i} collects temperature data from the device node s_i and calculates the average temperature for one minute on the temperature data ($i = 1, 2$). Each *t-aggregate* node f_{11i} supports one input port and one output port whose types are *temperature*. Then, the *t-aggregate* nodes f_{111} and f_{112} send the average *temperature* data to a parent fog node f_{11} supporting a process *t-merge*. The *t-merge* node f_{11} merges average temperature from child *t-aggregate* nodes for every one minute and sends a tuple $\langle \tau, t \rangle$ to a *join* fog node f_1 , where t shows the average temperature data of time τ . The *t-merge* node f_{11} has one input port whose type is *temperature* and one output port whose type is *time-temperature*. The *h-sensor* nodes s_3 and s_4 collect

humidity data every one second and send the collected data to parent edge nodes f_{121} and f_{122} , respectively, each of which supports a process *h-aggregate* which calculates on the average value for one minute. The *h-aggregate* nodes f_{121} and f_{122} send the average humidity data to a parent fog node f_{12} which supports a process *h-merge*. The *h-aggregate* node f_{12} has one *humidity* input port and one *humidity* output port. The *h-merge* node f_{12} collects the average humidity data from child *h-aggregate* nodes f_{121} and f_{122} for every one minute and sends a tuple $\langle \tau, h \rangle$ to a *join* fog node f_1 , where h is the average humidity of time τ . The *h-merge* node f_{12} has one *humidity* input port and one *time-humidity* output port. The *join* fog node f_1 receives the average values of temperature and humidity data from the child *t-merge* node f_{11} and *h-merge* node f_{12} , respectively, and joins the temperature data $\langle \tau, t \rangle$ and humidity data $\langle \tau, h \rangle$ for each time τ . The *join* node f_1 has two input ports, one for *time-temperature* type data and the other for *time-humidity* type data and one *time-temperature-humidity* output port. Then, the *join* node f_1 sends the *time-temperature-humidity* data $\langle \tau, t, h \rangle$ to a root *store* node f which is a server in a cloud. Every data from the *join* node f_1 is stored in the database of the *store* node f .

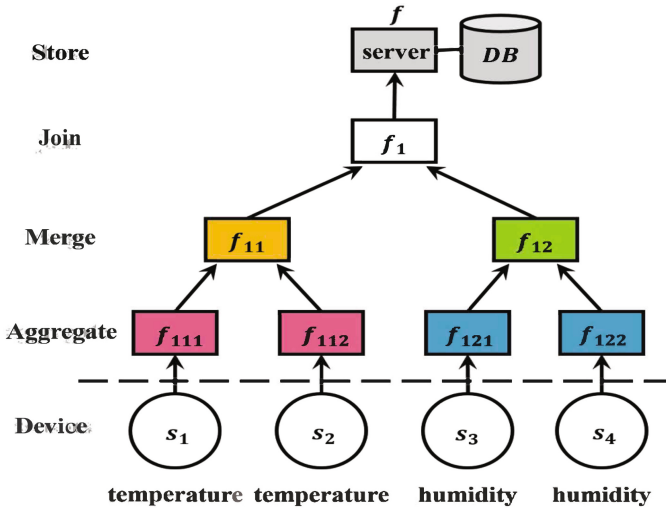


Fig. 2. Example of the TBFC model.

3 Power Consumption and Computation Models of the TBFC Model

Each fog node consumes electric energy [J] to communicate with parent and child nodes and to calculate output data on input data from child nodes. In

our previous studies, the *SPC* (Simple Power Consumption) model [7,8] and the *MLPCM* (Multi-Level Power Consumption) model [16–19] are proposed as power consumption models of a computer to perform application processes. In this paper, we consider the SPC model to obtain energy consumption of a fog node since a small computer like Raspberry Pi3 [34] is used to realize a fog node which follows the SPC model. In addition, more processes than the number of threads are usually performed on each server. Here, the server follows the SPC model. The power consumption NE_I [W] of a fog node f_I is given as follows:

[SPC model]

$$NE_I = \begin{cases} \max E_I & \text{if at least one process is active on } f_I. \\ \min E_I & \text{otherwise.} \end{cases} \quad (1)$$

For example, for a Raspberry Pi3 node f_I [34], $\max E_I$ and $\min E_I$ are 3.7 and 2.1 [W], respectively. On the other hand, $\max E_I$ and $\min E_I$ are 301.3 and 126.1 [W] for a server HP DL360 [6], respectively.

A node f_I consumes power to receive and send data. In this paper, we assume a node f_I consumes the power RE_I and SE_I [W] to receive and send data, respectively. $RE_I = re_I \cdot \max E_I$ and $SE_I = se_I \cdot \max E_I$ where $re_I (\leq 1)$ and $se_I (\leq 1)$ are constants. For a Raspberry Pi3 node f_I , $se_I = 0.68$ and $re_I = 0.73$ [31].

Next, we discuss the execution time [sec] of a process p_I of a node f_I . In this paper, we assume the execution time $ET_I(x)$ [sec] of a process p_I to calculate on input data of size x is $O(x)$ or $O(x^2)$ as discussed in papers [28–30,32], i.e. $ET_I(x)$ [sec] is $cc_I \cdot x$ or $cc_I \cdot x^2$ where cc_I is a constant. A process p_I is typed $O1$ and $O2$ iff $ET_I(x)$ is $cc_I \cdot x$ and $cc_I \cdot x^2$, respectively. The size $|od_I|$ of the output data od_I is $rr_I \cdot x$ for the size x of the input data. Here, rr_I is a *reduction* ratio of the fog node f_I . For example, if a fog node f_I obtains an average value od_I of m pieces of input data, the reduction ratio rr_I is $1/m (\leq 1)$.

A node f_I receives the input data ID_I from the child nodes and sends the output data od_I to the parent node. In this paper, it takes time $RT_I(x)$ and $ST_I(x)$ [sec] to receive and send data of size x , respectively, i.e. $RT_I(x) = rc_I \cdot x$ and $ST_I(x) = sc_I \cdot x$ where rc_I and sc_I are constants. $sc_I/rc_I = 0.22$ for a Raspberry Pi3 node f_I .

Suppose a node f_I receives the input data ID_I of size x from the child nodes. It totally takes time $TT_I(x) = RT_I(x) + ET_I(x) + ST_I(rr_I \cdot x)$ [sec]. The node f_I totally consumes the energy $TE_I(x) = RT_I(x) \cdot RE_I + ET_I(x) \cdot NE_t + ST_I(rr_I \cdot x) \cdot SE_I$ [W sec (J)].

In this paper, we assume the computation rate CR of a root node f is one. We measure the execution time of a process on a server [6] and a fog node f_I of Raspberry PI [34]. The computation rate CR_I of the fog node f_I is 0.18.

4 Evaluation

We consider a balanced tree T of fog nodes of the TBFC model in the evaluation. A TBFC tree T is specified as $\langle b, h \rangle$ -tree where b is the breadth of the tree T ,

i.e. the number of child nodes of each node and h is the height of the tree T . A root node f indicates a server in a cloud. Each node f_I has b (> 0) child nodes $f_{Ii_1}, \dots, f_{Ii_b}$ where $I = \langle i_1 \dots i_l \rangle$ ($1 \leq i_j \leq b, j = 1, \dots, l, l < h$). Here, $l + 1$ ($= |I| + 1$) shows a level of a fog node f_I in the tree T . Each node f_I supports a process p_I to calculate the output data od_I on the input data $ID_I = \{id_{Iij} | i = 1, \dots, b\}$ from child nodes $f_{Ii_1}, \dots, f_{Ii_b}$ and then sends the output data od_I to the parent node f_I as presented in the TBFC model. Every leaf node $f_{i_1 \dots i_{h-1}}$ is at the same level h and named an edge node. Every edge node $f_{i_1 \dots i_{h-1}}$ receives sensor data from the device node $s_{i_1 \dots i_{h-1}}$ at the same time every its [sec].

Table 1 shows the computation rate CR_I and the maximum power $maxE_I$ and minimum power $minE_I$ of a node f_I . The computation rate CR of a root node f is assumed to be one. The computation rate CR_I of a fog node f_I is 18 [%] of the root node f , which is obtained through our experiment [29, 31].

Table 1. Parameters of nodes

node f_I	CR_I	$minE_I$ [W]	$maxE_I$ [W]
root (server)	1.0	126.1	301.3
fog (Raspberry pi)	0.18	2.1	3.7

In the evaluation, we consider a $\langle b, 3 \rangle$ -tree T , i.e. the height h of the tree T is three ($h = 3$) and each node has b (> 0) child nodes. The tree T is composed of one root node, six fog nodes, and four device nodes. Each node f_I supports a process p_I to calculate the output data od_I on the input data ID_I sent from child nodes. In the evaluation, every node in the tree T supports an $O1$ or $O2$ type of process. We assume the constant cc_I of the computation rate $CR_I = cc_I \cdot x$ for size x of input data is one for each node f_I ($cc_I = 1$). The reduction ratio rr_I of each node f_I is 0.4. In the evaluation, we consider three processes p_1, p_2 , and p_3 , which are supported by a root node f , a fog node f_i , and an edge node f_{ij} ($i = 1, \dots, b, j = 1, \dots, b$), respectively. The process p_3 receives sensor data from device nodes and returns the output data to the process p_2 . Then, the output data of the process p_2 is sent to the process p_1 . Each node f_I has a pair of a conjunctive input port and an output port.

Each device node s_I sends sensor data to the edge node f_I every ist [sec]. In the evaluation, the inter-sensing time ist is four [sec]. Device nodes totally send sensor data of size tsd [B(byte)] to the edge nodes every ist [sec]. If there are m ($= b^{h-1}$) edge nodes in the tree T , each edge node f_I receives sensor data of size tsd/m [B] from a device node s_I .

In this paper, the receiving time $RT_I(x)$ and sending time $ST_I(x)$ [sec] to receive and send data of size x are assumed to be zero. This means, each node is assumed not to consume energy to send and receive data.

The delivery time DT [sec] of each sensor data is time from a device node sends sensor data to an edge node until a root node finishes processing the sensor

data. In the simulation, we evaluate the TBFC $\langle b, h \rangle$ -tree in terms of the total energy consumption TE [W sec] of the nodes and the delivery time DT [sec] of each sensor data.

Figure 3 shows the total energy consumption TE [W sec] of the nodes for breadth b of the $\langle b, 3 \rangle$ -tree T . Here, $O1$ and $O2$ show that processes are $O1$ and $O2$ types, respectively. The inter-sensing time (ist) is four [sec]. “ $b = 0$ ” means the CC model, i.e. every sensor data is sent to a server in the cloud and every process is performed on a server in the cloud. As shown in Fig. 3, the total energy consumption TE of nodes in the TBFC model is smaller than the CC model. For example, the TE of the TBFC model is 25 [%] and 16 [%] of the CC model for $O1$ and $O2$ process types, respectively. In addition, the more number of child nodes of each node, the smaller energy TE is consumed.

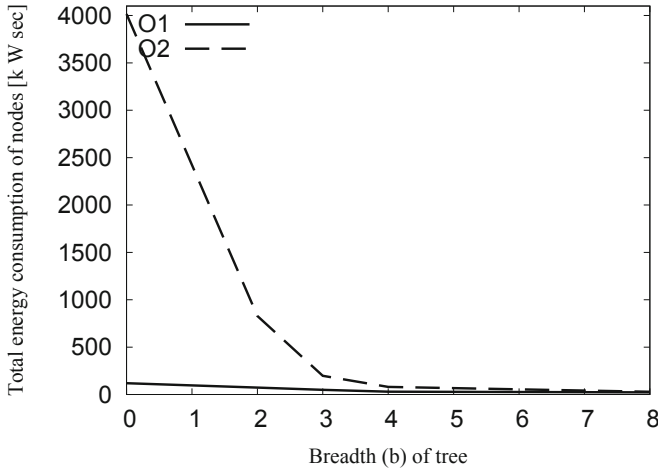


Fig. 3. Total energy consumption TE of nodes.

Figure 4 shows the delivery time DT [sec] of each sensor data, i.e. how long it takes to deliver sensor data to the root node for breadth b of the tree. The inter-sensing time (ist) is four [sec]. “ $b = 0$ ” stands for the CC model. The DT of the TBFC model is shorter than the CC model. For example, the DT of the TBFC model is 32 [%] and 7 [%] of the CC model for $O1$ and $O2$ process types, respectively. This means, a root node can make a decision on actions to be performed on devices earlier than the CC model. The more number of child nodes of each node, the shorter delivery time DT .

Figure 5 shows the average queue length QL [/sec], i.e. number of input data in the receipt queue of edge nodes of level 3 for inter-sensing time ist . In the evaluation, the QL of every other non-edge node is zero. If each node f_I receives input data from child nodes before sending the current output data to the parent node, the input data waits in the receipt queue of the node f_I . In the evaluation, sensor data arrives at each edge node f_I before finishing processing the preceding

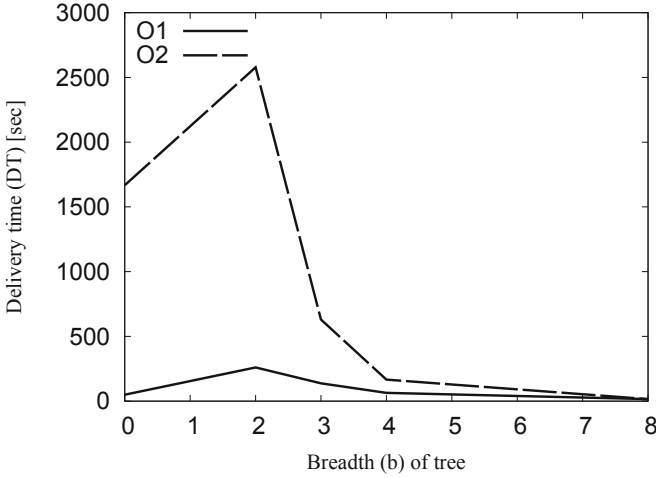


Fig. 4. Delivery time DT of sensor data.

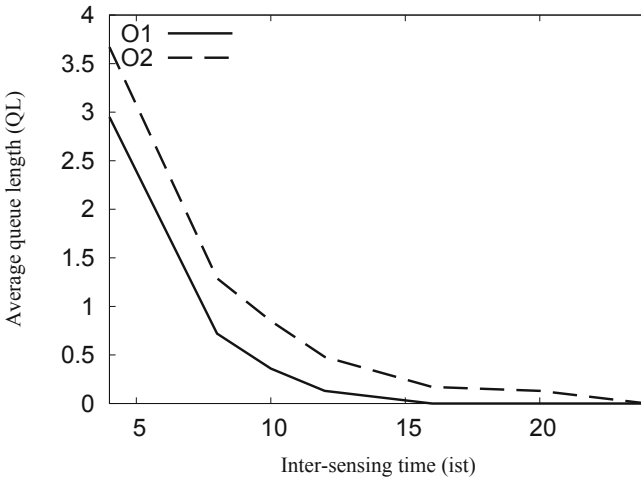


Fig. 5. Average queue length QL .

sensor data. The shorter inter-sensing time ist , the longer the queue length QL is. The longer queue length QL of a node, the longer time to deliver data to the root node. If additional nodes are deployed at level where the queue length of nodes is longer, the delivery time DT of sensor data can be reduced.

5 Concluding Remarks

It is critical to reduce the energy consumption of the IoT to realize green societies. The FC model is discussed to efficiently realize the IoT. The TBFC model

is a tree-structured model of fog nodes to reduce the total energy consumption of the nodes. In this paper, we evaluated the TBFC model compared with the CC model where device nodes periodically send sensor data to edge nodes. We showed the total energy consumption and delivery time of the TBFC model are shorter than the CC model.

As on-going studies, we are now discussing how to dynamically change the tree structure of the TBFC model so that input data does not wait at every fog node.

Acknowledgment. This work is supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 22K12018.

References

1. Dayarathna, M., Wen, Y., Fan, R.: Data center energy consumption modeling: a survey. *IEEE Commun. Surv. Tutor.* **18**(1), 732–787 (2016)
2. Natural Resources Defense Council: Data center efficiency assessment - scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers (2014). <http://www.nrdc.org/energy/files/data-center-efficiency-assessment-IP.pdf>
3. Qian, L., Luo, Z., Du, Y., Guo, L.: Cloud computing: an overview. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) *CloudCom 2009*. LNCS, vol. 5931, pp. 626–631. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10665-1_63
4. Hanes, D., Salgueiro, G., Grossetete, P., Barton, R., Henry, J.: *IoT Fundamentals: Networking Technologies, Protocols, and Use Cases for the Internet of Things*, 1st edn., p. 576. Cisco Press (2017)
5. Rahmani, A.M., Liljeborg, P., Preden, J.-S., Jantsch, A.: *Fog Computing in the Internet of Things*, 1st edn., p. 172. Springer, Heidelberg (2018)
6. HPE: HP server DL360 Gen 9. <https://www.techbuyer.com/cto/servers/hpe-proliant-dl360-gen9-server>
7. Enokido, T., Aikebaier, A., Takizawa, M.: Process allocation algorithms for saving power consumption in peer-to-peer systems. *IEEE Trans. Ind. Electron.* **58**(6), 2097–2105 (2011)
8. Enokido, T., Aikebaier, A., Takizawa, M.: A model for reducing power consumption in peer-to-peer systems. *IEEE Syst. J.* **4**(2), 221–229 (2010)
9. Enokido, T., Aikebaier, A., Takizawa, M.: An extended simple power consumption model for selecting a server to perform computation type processes in digital ecosystems. *IEEE Trans. Ind. Inform.* **10**(2), 1627–1636 (2014)
10. Enokido, T., Takizawa, M.: Integrated power consumption model for distributed systems. *IEEE Trans. Ind. Electron.* **60**(2), 824–836 (2013)
11. Enokido, T., Duolikun, D., Takizawa, M.: The energy consumption laxity-based algorithm to perform computation processes in virtual machine environments. *Int. J. Grid Util. Comput.* **10**(5), 545–555 (2019)
12. Enokido, T., Duolikun, D., Takizawa, M.: The improved redundant active time-based (IRATB) algorithm for process replication. In: Barolli, L., Woungang, I., Enokido, T. (eds.) *AINA 2021*. LNNS, vol. 225, pp. 172–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-75100-5_16

13. Enokido, T., Duolikun, D., Takizawa, M.: The redundant active time-based algorithm with forcing meaningless replica to terminate. In: Barolli, L., Yim, K., Enokido, T. (eds.) CISIS 2021. LNNS, vol. 278, pp. 206–213. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79725-6_20
14. Enokido, T., Duolikun, D., Takizawa, M.: The improved redundant active time-based algorithm with forcing termination of meaningless replicas in virtual machine environments. In: Barolli, L., Chen, H.-C., Enokido, T. (eds.) NBIS 2021. LNNS, vol. 313, pp. 50–58. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-84913-9_5
15. Kataoka, H., Duolikun, D., Sawada, A., Enokido, T., Takizawa, M.: Energy-aware server selection algorithms in a scalable cluster. In: Proceedings of the 30th International Conference on Advanced Information Networking and Applications, pp. 565–572 (2016)
16. Kataoka, H., Sawada, A., Dilawaer, D., Enokido, T., Takizawa, M.: Multi-level power consumption and computation models and energy-efficient server selection algorithms in a scalable cluster. In: Proceedings of the 19th International Conference on Network-Based Information Systems, pp. 210–217 (2016)
17. Kataoka, H., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: Multi-level power consumption model and energy-aware server selection algorithm. *Int. J. Grid Util. Comput.* **8**(3), 201–210 (2017)
18. Duolikun, D., Enokido, T., Takizawa, M.: Energy-efficient dynamic clusters of servers. In: Proceedings of the 8th International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 253–260 (2013)
19. Duolikun, D., Enokido, T., Takizawa, M.: Static and dynamic group migration algorithms of virtual machines to reduce energy consumption of a server cluster. In: Nguyen, N.T., Kowalczyk, R., Xhafa, F. (eds.) Transactions on Computational Collective Intelligence XXXIII. LNCS, vol. 11610, pp. 144–166. Springer, Heidelberg (2019). https://doi.org/10.1007/978-3-662-59540-4_8
20. Duolikun, D., Enokido, T., Takizawa, M.: Simple algorithms for selecting an energy-efficient server in a cluster of servers. *Int. J. Commun. Netw. Distrib. Syst.* **21**(1), 1–25 (2018)
21. Duolikun, D., Watanabe, R., Enokido, T., Takizawa, M.: An eco migration algorithm of virtual machines in a server cluster. In: Proceedings of IEEE the 32nd International Conference on Advanced Information Networking and Applications, pp. 189–196 (2018)
22. Duolikun, D., Enokido, T., Takizawa, M.: Energy-efficient group migration of virtual machines in a cluster. In: Barolli, L., Takizawa, M., Xhafa, F., Enokido, T. (eds.) AINA 2019. AISC, vol. 926, pp. 144–155. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-15032-7_12
23. Duolikun, D., Enokido, T., Barolli, L., Takizawa, M.: A monotonically increasing (MI) algorithm to estimate energy consumption and execution time of processes on a server. In: Barolli, L., Chen, H.-C., Enokido, T. (eds.) NBIS 2021. LNNS, vol. 313, pp. 1–12. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-84913-9_1
24. Duolikun, D., Enokido, T., Barolli, L., Takizawa, M.: An energy-efficient algorithm to make virtual machines migrate in a server cluster. In: Barolli, L., Kulla, E., Ikeda, M. (eds.) EIDWT 2022. LNDECT, vol. 118, pp. 130–141. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95903-6_15
25. Duolikun, D., Enokido, T., Barolli, L., Takizawa, M.: An energy consumption model of servers to make virtual machines migrate. In: Barolli, L., Hussain, F.,

- Enokido, T. (eds.) AINA 2022. LNNS, vol. 449, pp. 24–36. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99584-3_3
26. Duolikun, D., Enokido, T., Barolli, L., Takizawa, M.: Autonomous migration of virtual machines to reduce energy consumption of servers. In: Barolli, L. (ed.) IMIS 2022. LNNS, vol. 496, pp. 18–30. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08819-3_3
 27. Mukae, K., Saito, T., Nakamura, S., Enokido, T., Takizawa, M.: Design and implementing of the dynamic tree-based fog computing (DTBFC) model to realize the energy-efficient IoT. In: Barolli, L., Natwichai, J., Enokido, T. (eds.) EIDWT 2021. LNDECT, vol. 65, pp. 71–81. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70639-5_7
 28. Oma, R., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: An energy-efficient model of fog and device nodes in IoT. In: Proceedings of the 32nd International Conference on Advanced Information Networking and Applications Workshops, pp. 301–306 (2018)
 29. Oma, R., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: An energy-efficient model for fog computing in the Internet of Things (IoT). *Internet of Things* **1–2**, 14–26 (2018)
 30. Oma, R., Nakamura, S., Enokido, T., Takizawa, M.: A tree-based model of energy-efficient fog computing systems in IoT. In: Barolli, L., Javaid, N., Ikeda, M., Takizawa, M. (eds.) CISIS 2018. AISC, vol. 772, pp. 991–1001. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-93659-8_92
 31. Oma, R., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: Evaluation of an energy-efficient tree-based model of fog computing. In: Barolli, L., Kryvinska, N., Enokido, T., Takizawa, M. (eds.) NBiS 2018. LNDECT, vol. 22, pp. 99–109. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-98530-5_9
 32. Oma, R., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: A fault-tolerant tree-based fog computing model. *Int. J. Web Grid Serv.* **15**(3), 219–239 (2019)
 33. Chida, R., et al.: Implementation of fog nodes in the tree-based fog computing (TBFC) model of the IoT. In: Barolli, L., Xhafa, F., Khan, Z.A., Odhabi, H. (eds.) EIDWT 2019. LNDECT, vol. 29, pp. 92–102. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12839-5_8
 34. Raspberry Pi 3 Model B (2016). <https://www.raspberrypi.org/products/raspberrypi-3-model-b>



Evaluation of the Information Flow Control in the Fog Computing Model

Shigenari Nakamura¹(✉), Tomoya Enokido², and Makoto Takizawa³

¹ Tokyo Metropolitan Industrial Technology Research Institute, Tokyo, Japan

`nakamura.shigenari@iri-tokyo.jp`

² Rissho University, Tokyo, Japan

`eno@ris.ac.jp`

³ Hosei University, Tokyo, Japan

`makoto.takizawa@computer.org`

Abstract. In the IoT (Internet of Things), data are exchanged among subjects and objects in devices through manipulating objects. Even if subjects manipulate objects in accordance with the CBAC (Capability-Based Access Control) model, the subjects can get data which are not allowed to be gotten by the subjects, i.e. illegal information flow and late information flow occur. Hence, the OI (Operation Interruption) and TBOI (Time-Based OI) protocols where operations occurring illegal and late types of information flows are interrupted are implemented. Moreover, capability token selection algorithms are proposed and applied to the protocols. The protocols are implemented and evaluated in terms of the request processing time, communication traffic, and electric energy consumption. However, the more number of operations are interrupted to prevent both types of illegal and late information flows because the amount of data kept by entities monotonically increases through manipulating objects in the protocols. Therefore, reduction of the number of operations interrupted is important. For this aim, an FC (Fog Computing) model of the IoT where data from devices are processed in a fog layer and the processed data are sent to subjects is considered in this paper. In the evaluation, it is shown that the number of operations interrupted is reduced in the FC-based protocols compared with the conventional protocols.

Keywords: IoT (Internet of Things) · Device security · CBAC (Capability-Based Access Control) model · Information flow control · FC (Fog Computing) model

1 Introduction

It is widely recognized that access control models [3] are useful to make information systems secure. For the IoT (Internet of Things) [17], the CBAC (Capability-Based Access Control) model [5] is considered where capability tokens which are collections of access rights are issued to subjects. Only the authorized subjects

can manipulate objects in devices only in the authorized operations. Through manipulating objects in devices, data are exchanged among subjects and objects. Here, subjects might get data via other subjects and objects even if the subjects are granted no access right to get the data, i.e. illegal information flow might occur [8–11]. Moreover, a subject might get data generated out of the validity period of a capability token to get the data. Even if the time τ is not within the validity period of the capability token, the subject sb_i can get the data generated at time τ . Here, the data are older than the subject sb_i expects to get, i.e. the data come to the subject sb_i late [12].

In order to solve both types of illegal and late information flow problems in the IoT, the OI (Operation Interruption) [11] and TBOI (Time-Based OI) [12] protocols are implemented in a Raspberry Pi3 Model B+ [1] with Raspbian [2] which is regarded as an IoT device. A communication protocol between subjects and devices is the CoAP (Constrained Application Protocol) [19], which is implemented in CoAPthon3 [21]. In the OI and TBOI protocols, operations occurring illegal information flow and both types of illegal and late information flows are interrupted, i.e. not performed, at devices. In the evaluation, the request processing time gets longer as the number of capability tokens whose signatures are verified in devices increases in these protocols.

In order to reduce the number of capability tokens verified, a pair of algorithms, the MRCTSD (Minimum Required Capability Token Selection for Devices) [14] and MRCTSS (MRCTS for Subjects) [13] algorithms, are proposed and applied to the OI and TBOI protocols. In the MRCTSD algorithm, the request processing time is shortened because only the minimum required capability tokens are selected and used to make authorization decisions. In the MRCTSS algorithm, the communication traffic among subjects and objects is reduced because unnecessary capability tokens are not sent from subjects to devices.

In our previous studies [15], an electric energy consumption model of a Raspberry Pi 3 Model B+ equipped with Raspbian supporting the protocols is proposed. In the protocols, an authorization process is modeled to be a computation process which uses CPU resources like scientific computation. In the MLPC (Multi-Level Power Consumption) and MLC (ML Computation) models [7], devices are mainly characterized in terms of the numbers of cores and threads of a CPU. This means, the power consumption of a device depends on the numbers of active cores and threads. An energy consumption model of the Raspberry Pi supporting the protocols is proposed based on the MLPC and MLC models. Based on the models, electric energy consumption of each protocol is made clear in a simulation evaluation [16]. In the OI and TBOI protocols with capability token selection algorithms, the number of capability tokens used to make authorization decisions is reduced. Here, the request processing time is shortened compared with the conventional OI and TBOI protocols. Hence, the electric energy consumption is reduced compared with the conventional OI and TBOI protocols.

In the protocols, the amount of data kept by entities monotonically increases through manipulating objects. The more data are kept by entities, the more number of both types of illegal and late information flows occur. Therefore, the more number of operations are interrupted. In this paper, an FC (Fog Computing) model of the IoT [4] is considered to reduce the number of operations interrupted. In the FC model, a fog layer composed of fog nodes is introduced between devices and subjects. Data from devices are processed in the fog layer and the processed data are sent to subjects. Here, since the amount of data exchanged among entities is reduced, the number of both types of illegal and late information flows is also reduced. In the evaluation, the FC-based protocols are evaluated in terms of the number of operations interrupted. It is shown that the number of operations interrupted is reduced in the FC-based protocols compared with the conventional protocols.

In Sect. 2, the system model and types of information flow relations are discussed. In Sect. 3, the information flow control in the FC model to prevent both types of information flows is discussed. In Sect. 4, the FC-based protocols are evaluated in terms of the number of operations interrupted.

2 System Model

2.1 CBAC (Capability-Based Access Control) Model

In an IoT, there are the numbers dn and sbn of devices d_1, \dots, d_{dn} ($dn \geq 1$) and subjects sb_1, \dots, sb_{sbn} ($sbn \geq 1$), respectively. Each device d_k holds the number on^k of objects $o_1^k, \dots, o_{on^k}^k$ ($on^k \geq 1$). A term, “object o_m^k ” stands for a component object in the device d_k . There are various types of devices such as sensors, actuators, and hybrid devices. A sensor d_k collects data obtained by sensing events occurring in physical environment and stores the data in its object o_m^k . An actuator d_k receives data collected by sensors and stores the data in its object o_m^k . The actuator d_k performs actions on the physical environment based on the data. A hybrid device d_k is equipped with both the sensors and actuators.

Subjects manipulate data of objects in devices. In order to make the devices secure, a CBAC model [5] is considered in the IoT. Data collected by sensors are stored at an object in the sensors. A subject gets the data by accessing the object. The subject designates actions for actuators based on the data. Here, the data are put to objects of the actuators by the subject. Hence, data are got and put from and to hybrid devices by subjects. Each subject sb_i is issued a set CAP^i which consists of the number cn^i of capability tokens $cap_1^i, \dots, cap_{cn^i}^i$ ($cn^i \geq 1$).

A capability token cap_g^i is designed as shown in the papers [11,12]. Let $cap_g^i.IS$ and $cap_g^i.SU$ be public keys of an issuer and a subject of the capability token cap_g^i , respectively. $cap_g^i.SG$ is a signature generated with the private key of the issuer. These signatures and keys are generated in the ECDSA (Elliptic Curve Digital Signature Algorithm) [6] and then encoded into Base64 form.

The capability token cap_g^i indicates how the subject sb_i can manipulate objects in a device shown in $cap_g^i.DE$. Access rights field of the capability token cap_g^i indicates object shown in $cap_g^i.OB$ can be manipulated in operation shown in $cap_g^i.OP$. The capability token cap_g^i is valid at time τ where $cap_g^i.NB < \tau < cap_g^i.NA$. Capability tokens are included in the payload field of a CoAP request.

Let dt_m^k be data of an object o_m^k . If a subject sb_i tries to manipulate the data dt_m^k of the object o_m^k in a device d_k in an operation op , the subject sb_i sends an access request with a capability token cap_g^i to specify the subject sb_i is allowed to manipulate the object o_m^k in the operation op to the device d_k . If the device d_k confirms that the subject sb_i is allowed to manipulate the object o_m^k in the operation op , the operation op is performed on the object o_m^k . Otherwise, the access request is rejected. Since the device d_k just checks the capability token cap_g^i to authorize the subject sb_i , it is easier to adopt the CBAC model to the IoT than the ACL (Access Control List)-based models such as RBAC (Role-Based Access Control) [18] and ABAC (Attribute-Based Access Control) [22] models.

Let a pair $\langle o, op \rangle$ be an access right. Subjects issued a capability token including an access right $\langle o, op \rangle$ is allowed to manipulate data of an object o in an operation op . A set of objects whose data a subject sb_i is allowed to get is $IN(sb_i)$ i.e. $IN(sb_i) = \{o_m^k \mid \langle o_m^k, get \rangle \in cap_g^i \wedge cap_g^i \in CAP^i\}$.

Through manipulating data of objects in devices, the data are exchanged among subjects and objects. Objects whose data flow into entities are referred to as *source* objects for these entities. Let $o_m^k.sO$ and $sb_i.sO$ are sets of *source* objects of an object o_m^k and a subject sb_i , respectively, which are initially ϕ .

A capability token cap_g^i has the validity period. Let a pair of times $gt^i.st(o_m^k)$ and $gt^i.et(o_m^k)$ be the start and end time when a subject sb_i is allowed to get data dt_m^k from the object o_m^k . The time when data dt_m^k of an object o_m^k are generated is referred to a generation time. Let $minOT_m^k(o_n^l)$ and $minSBT^i(o_n^l)$ be the earliest generation times of data dt_n^l of an object o_n^l which flow to an object o_m^k and a subject sb_i , respectively.

2.2 Information Flow Relations

If a subject sb_i issues a get operation to get data dt_m^k from an object o_m^k , information flow from the object o_m^k to the subject sb_i occurs. Based on the CBAC model, types of information flow relations on objects and subjects are defined as follows:

Definition 1. An object o_m^k *flows* to a subject sb_i ($o_m^k \rightarrow sb_i$) iff (if and only if) $o_m^k.sO \neq \phi$ and $o_m^k \in IN(sb_i)$.

Definition 2. An object o_m^k *legally flows* to a subject sb_i ($o_m^k \Rightarrow sb_i$) iff $o_m^k \rightarrow sb_i$ and $o_m^k.sO \subseteq IN(sb_i)$.

Definition 3. An object o_m^k *illegally flows* to a subject sb_i ($o_m^k \mapsto sb_i$) iff $o_m^k \rightarrow sb_i$ and $o_m^k.sO \not\subseteq IN(sb_i)$.

Definition 4. An object o_m^k timely flows to a subject sb_i ($o_m^k \Rightarrow_t sb_i$) iff $o_m^k \Rightarrow sb_i$ and $\forall o_n^l \in o_m^k \cdot sO (gt^i \cdot st(o_n^l) \leq \min OT_m^k(o_n^l) \leq gt^i \cdot et(o_n^l))$.

Definition 5. An object o_m^k flows late to a subject sb_i ($o_m^k \mapsto_l sb_i$) iff $o_m^k \Rightarrow sb_i$ and $\exists o_n^l \in o_m^k \cdot sO \neg(gt^i \cdot st(o_n^l) \leq \min OT_m^k(o_n^l) \leq gt^i \cdot et(o_n^l))$.

2.3 FC (Fog Computing) Model

In addition to computers, a huge number and various types of devices like sensors and actuators are interconnected in the IoT (Internet of Things) [4, 17]. Here, a large volume of sensor data are transmitted from sensors to subjects in networks. Subjects decide actions by analyzing sensor data and send the actions to actuators. On receipt of an action, each actuator performs the action on physical environment. In order to reduce the network traffic and satisfy the time constraints between sensors and actuators, a fog layer is introduced between devices and subjects as shown in Fig. 1.

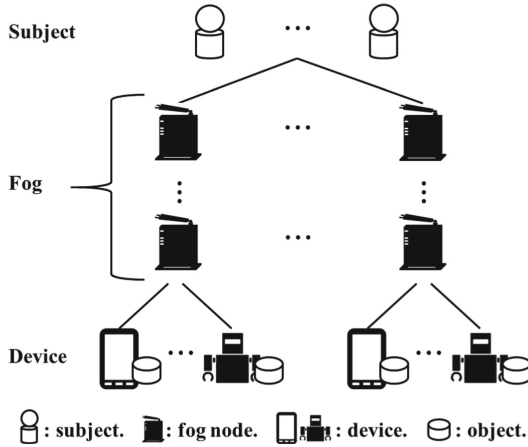


Fig. 1. FC model.

The device layer is composed of devices such as sensors, actuators, and hybrid devices. The fog layer is composed of fog nodes [4]. Fog nodes are interconnected with other fog nodes in networks. Fog nodes support the routing function where messages are routed to destination nodes, i.e. routing between subjects and devices like network routers. Thus, fog nodes receive data and forward the data to subjects in fog-to-fog communication.

More importantly, a fog node does some computation on a collection of input data from sensors and other fog nodes. The input data are processed and new output data, i.e. processed data, are generated by a fog node. For example, an average value is calculated from a collection of data from sensors. Here, the

output data are smaller than the input data. Data processed by a fog node are sent to neighbor fog nodes and subjects finally receive the data processed by fog nodes.

Suppose a subject sb_i issues a *get* operation on an object o_m^k to a device d_k . The data dt_m^k of the object o_m^k arrive at the fog layer before being sent to the subject sb_i . Here, the data dt_m^k are processed and summarized data are generated by a set $F_m^{k,i}$ of fog nodes for the subject sb_i . Let $F^i(o_m^k.sO)$ be a set of *source* objects whose data are included in the summarized data. The set $F^i(o_m^k.sO)$ is decided in accordance with the data processing of the fog layer. Finally, data of objects in $F^i(o_m^k.sO)$ flow to the subject sb_i .

3 Information Flow Control

3.1 Protocols

In the CBAC model, data are exchanged among subjects and objects. Here, a subject sb_i may get data dt_m^k of an object o_m^k flowing to another object o_n^l by accessing the object o_n^l even if the subject sb_i is not allowed to get the data dt_m^k from the object o_m^k , i.e. illegal information flow occurs. In addition, a subject sb_i may get data dt_m^k from an object o_m^k generated out of validity period of a capability token cap_g^i to get the data dt_m^k . Here, the data dt_m^k are older than the subject sb_i expects to get, i.e. information comes to the subject sb_i *late*. In order to prevent illegal information flow and both illegal and late types of information flows, the OI (Operation Interruption) [11] and TBOI (Time-Based OI) [12] protocols are implemented, respectively. In the implementation, a Raspberry Pi 3 Model B+ [1] equipped with Raspbian [2] is used as a device. A communication protocol between a subject and the device is the CoAP [19], which is implemented in CoAPthon3 [20].

In order to prevent the illegal information flow, sets of *source* objects are manipulated in the OI protocol. Here, if data dt_m^k of an object o_m^k flow to an entity, the object o_m^k is added to a *source* object set of the entity. For example, since data flow from an object o_m^k to a subject sb_i in a *get* operation, the set $o_m^k.sO$ of the object o_m^k are added to the set $sb_i.sO$ of the subject sb_i . On the other hand, since data flow from a subject sb_i to an object o_m^k in a *put* operation, the set $sb_i.sO$ are added to the set $o_m^k.sO$. In the TBOI protocol, the earliest generation time of data of every *source* object is also updated. Based on the sets of *source* objects and the earliest generation time of data, the illegal information flow and late information flow are detected. The OI and TBOI protocols perform as follows:

[OI protocol] A *get* operation on an object o_m^k issued by a subject sb_i is interrupted if $o_m^k \Rightarrow sb_i$ does not hold.

[TBOI protocol] A *get* operation on an object o_m^k issued by a subject sb_i is interrupted if $o_m^k \Rightarrow_l sb_i$ does not hold.

In the implementation evaluation, it is shown that the time to make an authorization decision increases as the number of capability tokens used in the authorization process increases. Hence, the MRCTSD (Minimum Required Capability Token Selection for Devices) algorithm is proposed for the OI and TBOI protocols to select only the capability tokens required to make the authorization decision at devices [14]. After that, the MRCTSS (MRCTS for Subjects) algorithm is proposed to select the capability tokens sent from subjects to devices [13]. Here, since unnecessary capability tokens are not sent from subjects to devices, the communication traffic is reduced.

In the paper [16], the OI and TBOI protocols with the capability token selection algorithms are evaluated in terms of the electric energy consumption. Here, it is shown that the electric energy consumed by devices are reduced by the capability token selection algorithms.

3.2 FC-Based Protocols

In the protocols discussed in previous Sects., the amount of data kept by entities monotonically increases through manipulating objects. The more data are kept by entities, the more number of both types of illegal and late information flows occur. Therefore, the more number of operations are interrupted. In order to reduce the number of operations interrupted, an FC (Fog Computing) model in the IoT [4] is considered in this paper.

In the FC model, there is a fog layer composed of fog nodes between devices and subjects. Suppose a subject sb_i issues a *get* operation on an object o_m^k to a device d_k . The data dt_m^k of the object o_m^k arrive at the fog layer before being sent to the subject sb_i . Here, the data dt_m^k are processed and summarized data are generated by a set $F_m^{k,i}$ of fog nodes for the subject sb_i . As a result, data of objects in $F^i(o_m^k.sO)$ are sent to the subject sb_i . Therefore, information flow relations in the FC model are newly defined with the set $F^i(o_m^k.sO)$ of *source* objects as follows:

Definition 6. $o_m^k \rightarrow^F sb_i$ iff $F^i(o_m^k.sO) \neq \phi$ and $o_m^k \in IN(sb_i)$.

Definition 7. $o_m^k \Rightarrow^F sb_i$ iff $o_m^k \rightarrow^F sb_i$ and $F^i(o_m^k.sO) \subseteq IN(sb_i)$.

Definition 8. $o_m^k \mapsto^F sb_i$ iff $o_m^k \rightarrow^F sb_i$ and $F^i(o_m^k.sO) \not\subseteq IN(sb_i)$.

Definition 9. $o_m^k \Rightarrow_t^F sb_i$ iff $o_m^k \Rightarrow^F sb_i$ and $\forall o_n^l \in F^i(o_m^k.sO)$ ($gt^i.st(o_n^l) \leq \min OT_m^k(o_n^l) \leq gt^i.et(o_n^l)$).

Definition 10. $o_m^k \mapsto_t^F sb_i$ iff $o_m^k \Rightarrow^F sb_i$ and $\exists o_n^l \in F^i(o_m^k.sO)$ ($gt^i.st(o_n^l) \leq \min OT_m^k(o_n^l) \leq gt^i.et(o_n^l)$).

Generally, data from objects are processed and the summarized data are generated by fog nodes. After that, the summarized data are sent to subjects. Here, since the amount of data exchanged among entities decreases, the amount of data kept by entities slightly increases compared with the conventional system

models. As a result, the number of both types of illegal and late information flows is reduced. Hence, the number of operations interrupted is also reduced in the protocols.

In this paper, calculable data are assumed to be exchanged among entities. For example, three typical calculations to extract the maximum, minimum, and average values are considered. The set $F^i(o_m^k.sO)$ which indicates *source* objects of summarized data is decided in accordance with these calculations. These typical calculations and the set $F^i(o_m^k.sO)$ of *source* objects are summarized as follows:

- Extraction of maximum value: $\{o_n^l \mid dt_n^l \text{ in } o_m^k \text{ include the maximum value}\}$.
- Extraction of minimum value: $\{o_n^l \mid dt_n^l \text{ in } o_m^k \text{ include the minimum value}\}$.
- Extraction of average value: $\{o_n^l \mid dt_n^l \text{ in } o_m^k \text{ include a value closest to the average value}\}$.

In this paper, the FCOI (FC-based OI) and FCTBOI (FC-based TBOI) protocols are proposed. In the FCOI protocol, the sets of *source* objects are updated as follows:

1. Initially, $sb_i.sO = o_m^k.sO = \phi$ for every subject sb_i and object o_m^k ;
2. If a device d_k generates data by sensing events occurring around itself and stores the data to its object o_m^k , $o_m^k.sO = o_m^k.sO \cup \{o_m^k\}$;
3. If a subject sb_i issues a *get* operation on an object o_m^k , $sb_i.sO = sb_i.sO \cup F^i(o_m^k.sO)$;
($F^i(o_m^k.sO)$ is decided in accordance with the data processing in a fog layer)
4. If a subject sb_i issues a *put* operation on an object o_m^k , $o_m^k.sO = o_m^k.sO \cup sb_i.sO$;

In a *get* operation, data dt_m^k from the object o_m^k are processed and summarized data are generated by fog nodes in $F_m^{k,i}$. Hence, the objects in the set $F^i(o_m^k.sO)$ are added to the set $sb_i.sO$ of the subject sb_i .

On the other hand, in the FCTBOI protocol, the earliest generation times of data of *source* objects are also updated as follows:

1. Initially, $sb_i.sO = o_m^k.sO = \phi$ for every subject sb_i and object o_m^k ;
2. If a device d_k generates data by sensing events occurring around itself and stores the data to its object o_m^k at time τ .
 - a. If $minOT_m^k(o_m^k) = \text{NULL}$, $minOT_m^k(o_m^k) = \tau$;
 - b. $o_m^k.sO = o_m^k.sO \cup \{o_m^k\}$;
3. If a subject sb_i issues a *get* operation on an object o_m^k .
 - a. For each object o_n^l such that $o_n^l \in (sb_i.sO \cap F^i(o_m^k.sO))$, $minSBT^i(o_n^l) = \min(minSBT^i(o_n^l), minOT_m^k(o_n^l))$;
 - b. For each object o_n^l such that $o_n^l \notin sb_i.sO$ but $o_n^l \in F^i(o_m^k.sO)$, $minSBT^i(o_n^l) = minOT_m^k(o_n^l)$;
 - c. $sb_i.sO = sb_i.sO \cup F^i(o_m^k.sO)$;
4. If a subject sb_i issues a *put* operation on an object o_m^k .

- a. For each object o_n^l such that $o_n^l \in (sb_i.sO \cap o_m^k.sO)$, $minOT_m^k(o_n^l) = min(minOT_m^k(o_n^l), minSBT^i(o_n^l))$;
- b. For each object o_n^l such that $o_n^l \notin o_m^k.sO$ but $o_n^l \in sb_i.sO$, $minOT_m^k(o_n^l) = minSBT^i(o_n^l)$;
- c. $o_m^k.sO = o_m^k.sO \cup sb_i.sO$;

Based on the sets of *source* objects and the earliest generation times of data, the illegal information flow and late information flow are detected. The FCOI and FCTBOI protocols perform as follows:

[FCOI protocol] A *get* operation on an object o_m^k issued by a subject sb_i is interrupted if $o_m^k \Rightarrow^F sb_i$ does not hold.

[FCTBOI protocol] A *get* operation on an object o_m^k issued by a subject sb_i is interrupted if $o_m^k \Rightarrow_t^F sb_i$ does not hold.

Since the number of *source* objects exchanged among entities are reduced by the processing of fog nodes, it is expected that the numbers of operations interrupted are reduced in the FCOI and FCTBOI protocols compared with the conventional OI and TBOI protocols, respectively.

4 Evaluation

The FCOI and FCTBOI protocols are evaluated in terms of the number of operations interrupted. In the evaluation, twenty devices and thirty subjects are considered ($dn = 20$, $sb_n = 30$). It is assumed that subjects issue only *get* and *put* operations. Three types of devices, sensors, actuators, and hybrid devices are considered. Sensors and actuators accept only *get* and *put* operations, respectively. On the other hand, hybrid devices accept both *get* and *put* operations.

Each device d_k obtains the number on^k of objects. The number on^k is randomly selected out of numbers 1, ..., 5. The type of each device d_k is randomly decided with the same probability. Initially, a set of *source* object $o_m^k.sO$ of each object o_m^k is empty.

Every subject sb_i is issued the number cn^i of capability tokens. Every capability token cap_g^i includes the number arn_g^i of access rights. Validity period vp_g^i of the capability token cap_g^i is randomly selected out of numbers 300, ..., 600 simulation steps. The numbers cn^i and arn_g^i are randomly selected out of numbers 5, ..., 15 and 1, ..., $on^k \cdot ops_n^k$, respectively. ops_n^k is the number of operations supported in the device d_k . Hence, if the device d_k is a sensor or actuator, $ops_n^k = 1$. Otherwise, i.e. the device d_k is a hybrid device, $ops_n^k = 2$. Access rights and a device d_k shown in $cap_g^i.DE$ in the capability token cap_g^i are randomly decided. The operation of an access right for an object o_m^k in a hybrid device d_k is decided to be *get* with probability 0.5. On the other hand, the operation is decided to be *put* with the same probability.

After the generation of devices and subjects, the following procedures are performed in every protocol:

1. Every sensor and hybrid device d_k collects data by sensing events with probability 0.5. Here, the sensing type is randomly selected so that the sensing is full one with probability 0.2. If the device d_k collects data, the data are stored in an object o_m^k randomly selected.
2. For every subject sb_i , the validity period vp_g^i of every capability token cap_g^i is decremented by one. If the validity period vp_g^i gets 0, the capability token cap_g^i is revoked from the subject sb_i .
3. Every subject sb_i which has no capability token is issued capability tokens randomly generated.
4. Every subject sb_i issues an operation with probability 0.7. If a subject sb_i decides to issue an operation, one access right obtained by the subject sb_i is randomly selected and the subject sb_i issues an operation according to the access right. In a *get* operation, the subject sb_i requests summarized data with probability 0.5. Here, data processed in a fog layer flow to the subject sb_i . Otherwise, raw data flow to the subject sb_i . On the other hand, in a *put* operation, data in the subject sb_i flow to the object. For each operation, full and partial types are randomly selected with probabilities 0.2 and 0.8, respectively.
5. In a *get* operation, data dt_m^k from an object o_m^k are processed at a fog layer for a subject sb_i . The calculation type is randomly decided with the same probability. The set $F^i(o_m^k.sO)$ of *source* objects is generated in accordance with the calculation and sent to the subject sb_i . After the processing, if $o_m^k \Rightarrow^F sb_i$ and $o_m^k \Rightarrow_t^F sb_i$ do not hold, the *get* operation is interrupted at the fog layer in the FCOI and FCTBOI protocols, respectively.

In the simulation, we make the following assumptions:

- If $o_m^k.sO = \phi$, a subject sb_i does not issue a *get* operation to a device d_k .
- If $sb_i.sO = \phi$, a subject sb_i does not issue a *put* operation to a device d_k .
- If $\exists o_n^l \in o_m^k.sO$ ($minOT_m^k(o_n^l) < minSBT^i(o_n^l)$), a subject sb_i does not issue a full *get* operation to a device d_k .
- If $\exists o_n^l \in sb_i.sO$ ($minSBT^i(o_n^l) < minOT_m^k(o_n^l)$), a subject sb_i does not issue a full *put* operation to a device d_k .

Let st be a simulation steps. The above procedures are assumed to be performed in one simulation step. This means, the above procedures are iterated st times. In the evaluation, one simulation step means one [s]. Here, $st = 0, 600, 1200, 1800, 2400, 3000, \text{ or } 3600$. The sets of devices and subjects are randomly generated twenty times. For a pair of given these sets, the above procedures for st are iterated ten times. Finally, the average number of operations interrupted is calculated.

Figure 2 shows the numbers of *get* operations interrupted in the OI, TBOI, FCOI, and FCTBOI protocols. In the FCOI and FCTBOI protocols, since data from objects are processed at fog nodes, the number of objects in the set of *source* objects are reduced. Here, the number of both types of illegal and late information flows are also reduced. Hence, the number of operations interrupted is reduced in the FCOI and FCTBOI protocols compared with the OI and TBOI protocols, respectively.

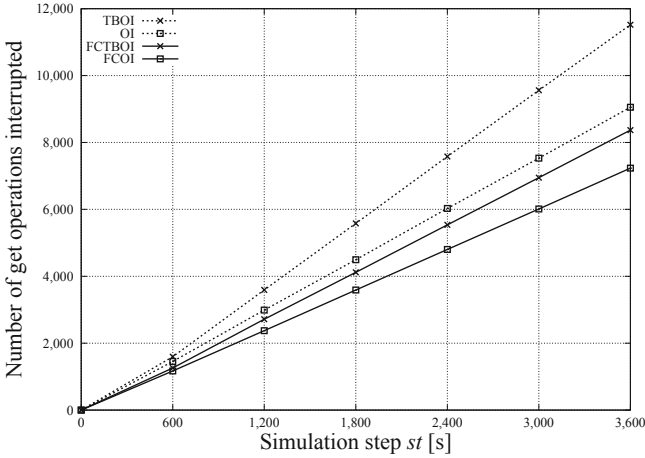


Fig. 2. Number of *get* operations interrupted.

5 Concluding Remarks

For the IoT (Internet of Things), the CBAC (Capability-Based Access Control) model was proposed where capability tokens which are collection of access rights are issued to subjects. Since data are exchanged among entities through manipulating objects, two types of illegal and late information flows occur. In order to prevent both types of illegal and late information flows from occurring, the OI (Operation Interruption) and TBOI (Time-Based OI) protocols were proposed. In addition, the MRCTSD (Minimum Required Capability Token Selection for Devices) and MRCTSS (MRCTS for Subjects) algorithms were proposed to make the OI and TBOI protocols more useful. In the protocols, the amount of data kept by entities monotonically increases through manipulating objects. As a result, the more number of operations are interrupted to prevent both types of illegal and late information flows. In order to reduce the number of operations interrupted, an FC (Fog Computing) model in the IoT is considered in this paper. In the FC model, data from devices are processed in a fog layer and the processed data are sent to subjects. Here, since the amount of data exchanged among entities is reduced, the number of both types of illegal and late information flows is also reduced. Hence, the number of operations interrupted is reduced in the FC-based protocols compared with the conventional protocols in the evaluation.

Acknowledgements. This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP22K12018.

References

1. Raspberry Pi 3 Model B+. <https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/>
2. Raspbian, version 10.3 (2020). <https://www.raspbian.org/>
3. Denning, D.E.R.: *Cryptography and Data Security*. Addison Wesley, Boston (1982)
4. Hanes, D., Salgueiro, G., Grossetete, P., Barton, R., Henry, J.: *IoT Fundamentals: Networking Technologies, Protocols, and Use Cases for the Internet of Things*. Cisco Press, Indianapolis (2018)
5. Hernández-Ramos, J.L., Jara, A.J., Marín, L., Skarmeta, A.F.: Distributed capability-based access control for the internet of things. *J. Internet Serv. Inf. Secur.* **3**(3/4), 1–16 (2013)
6. Johnson, D., Menezes, A., Vanstone, S.: The elliptic curve digital signature algorithm (ECDSA). *Int. J. Inf. Secur.* **1**(1), 36–63 (2001)
7. Kataoka, H., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: Multi-level power consumption model and energy-aware server selection algorithm. *Int. J. Grid Util. Comput.* **8**(3), 201–210 (2017)
8. Nakamura, S., Duolikun, D., Aikebaier, A., Enokido, T., Takizawa, M.: Read-write abortion (RWA) based synchronization protocols to prevent illegal information flow. In: *Proceedings of the 17th International Conference on Network-Based Information Systems*, pp. 120–127 (2014)
9. Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: A read-write abortion protocol to prevent illegal information flow in role-based access control systems. *Int. J. Space-Based Situated Comput.* **6**(1), 43–53 (2016)
10. Nakamura, S., Enokido, T., Takizawa, M.: Information flow control in object-based peer-to-peer publish/subscribe systems. *Concurr. Comput. Pract. Experience* **32**(8), e5118 (2020)
11. Nakamura, S., Enokido, T., Takizawa, M.: Implementation and evaluation of the information flow control for the internet of things. *Concurr. Comput. Pract. Experience* **33**(19), e6311 (2021)
12. Nakamura, S., Enokido, T., Takizawa, M.: Information flow control based on capability token validity for secure IoT: implementation and evaluation. *Internet of Things* **15**, 100,423 (2021)
13. Nakamura, S., Enokido, T., Takizawa, M.: Traffic reduction for information flow control in the IoT. In: Barolli, L. (ed.) *BWCCA 2021. LNNS*, vol. 346, pp. 67–77. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-90072-4_7
14. Nakamura, S., Enokido, T., Takizawa, M.: Capability token selection algorithms to implement lightweight protocols. *Internet of Things* **19**, 100,542 (2022)
15. Nakamura, S., Enokido, T., Takizawa, M.: Energy consumption model of a device supporting information flow control in the IoT. In: Barolli, L., Kulla, E., Ikeda, M. (eds.) *EIDWT 2022. LNDECT*, vol. 118, pp. 142–152. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95903-6_16
16. Nakamura, S., Enokido, T., Takizawa, M.: Energy consumption of the information flow control in the IoT: simulation evaluation. In: Barolli, L., Hussain, F., Enokido, T. (eds.) *AINA 2022. LNNS*, vol. 449, pp. 285–296. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99584-3_25
17. Oma, R., Nakamura, S., Duolikun, D., Enokido, T., Takizawa, M.: An energy-efficient model for fog computing in the internet of things (IoT). *Internet of Things* **1–2**, 14–26 (2018)

18. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-based access control models. *IEEE Comput.* **29**(2), 38–47 (1996)
19. Shelby, Z., Hartke, K., Bormann, C.: Constrained application protocol (CoAP). IETF Internet-draft (2013). <http://tools.ietf.org/html/draft-ietf-core-coap-18>
20. Tanganelli, G., Vallati, C., Mingozzi, E.: CoAPthon3 version 1.0.1 (2018). <https://github.com/Tanganelli/CoAPthon3>. Accessed 28 June 2020
21. Tanganelli, G., Vallati, C., Mingozzi, E.: CoAPthon: easy development of CoAP-based IoT applications with Python. In: *IEEE 2nd World Forum on Internet of Things (WF-IoT 2015)*, pp. 63–68 (2015)
22. Yuan, E., Tong, J.: Attributed based access control (ABAC) for web services. In: *Proceedings of the IEEE International Conference on Web Services (ICWS 2005)*, p. 569 (2005)



A Study of Network Attack Strategy Using AS Topology Map

Naoya Sekiguchi^(✉) and Hidema Tanaka

National Defense Academy, Yokosuka, Japan
{em61024, hidema}@nda.ac.jp

Abstract. The Internet is operated by interconnected networks of units called AS. In recent years, BGP hijackings have caused large-scale failures and interceptions. This paper focuses on the activities of it that targets AS and BGP. We analyze possible methods of it and propose a method to localize attack effectiveness. We derive a topology map of AS from BGP logs and analyze its characteristics of it. Focusing on strategies that change it and its characteristics, we assume two scenarios and three attack tactics. From our computer simulations, we can find the following two facts. First, if the adversary group wants to spread malware and disinformation, setting “fake ASs” is effective. Second, if the group wants to concentrate and confusion about information sharing, stopping some ASs is effective. These are easy to realize because the attacker can succeed only by rewriting ASPATH. On the other hand, as a countermeasure, we can find that setting a new AS can decrease such attack effectiveness.

1 Introduction

The objectives and techniques of network attacks are rapidly becoming more complex every day. In particular, it has been observed that attackers are organized into the adversary group. For example, a typical ATP (Advanced Threat Prevention), is organized and roles are divided among them. Also, in the Russian and Ukrainian wars, it involving civilians has been organized, and it has become necessary to take into account network attacks of a different scale and strategy than before. This paper considers the potential risks of the Internet structure for such a recent tendency.

The Internet is operated based on AS (Autonomous System). An AS is a set of networks managed following a unified policy. Each AS has its own “AS number” which is determined by IANA [4], RIRs, NIRs, and so on. To establish communication among ASs, AS exchange their route information “BGP (Border Gateway Protocol)” each other. BGP can control a huge number of routings. The administrator of each AS can connect to the Internet by configuring appropriate route advertisements. Therefore, incorrect route advertisements can cause serious failures in networks around the world. This incorrect route advertisement is called “BGP hijacking”. It has resulted in massive access disruptions and interceptions to corporate, financial and government institutions [1–3, 5, 8, 13]. Although countermeasures are being taken against it, they have not yet reached the practical security [6, 11, 12, 17]. It allows the adversary group to conduct large-scale disruption or interception without access to the target network or attacking the important servers which require high cost and techniques.

Analyzing the BGP logs allows us to derive a topology map of AS. The properties of the topology map can be estimated by the eigenvalues of the matrices derived from it. Such an analysis method is developed in the research field of network dynamics. Paper [19] has discussed network attack countermeasures that use these eigenvalues to select the most effective IP addresses and appropriate attack methods. Paper [19] derives a topology map of IP addresses in a specific area and changes their characteristics assuming Slow Read DDoS attacks. In this attack, the attackers are a small group and are not expected to be a sustained attack. However, if we assume the adversary group colludes with network companies or a hostile foreign party, a more extensive and sustained attack can be expected.

In this paper, we propose an attack strategy to evaluate possible attack tactics of the adversary group and to localize the damage by developing the scheme of previous studies. We focus on varying topology maps of AS and their characteristics, and we evaluate them based on two different scenarios and three different attack tactics. The effectiveness of the attack is evaluated by the increase of the eigenvalue. In this paper, we evaluate the feasibility of our proposed attack strategy by computer simulation using the database of the routeviews project [9], which collects from real communication. We demonstrate the effectiveness of our proposal using such an actual database, however, details are omitted because of ethical reasons.

2 Preliminaries

2.1 Outline

Network characteristics can be estimated using a topology map. An example of a previous study using eigenvalues of such a topology map is a chain bankruptcy analysis of banking transactions [7]. In this study, a topology map of banking transactions is derived and its eigenvalues are computed. By using these eigenvalues, they show that the failure of a megabank is not the only cause of a financial crisis. Network dynamics is a research field that analyzes phenomena using the characteristics of such networks.

Topology maps can be represented in several ways. In this paper, we use two types of integer matrices: adjacency matrices [15] and Laplacian matrices [20]. The eigenvalues of each matrix represent the characteristics of the topology map. In particular, we focus on two types of characteristics; “Spread of speed” and “Convergence”.

2.2 Adjacency Matrix

Let G be a topology map with n nodes. G can be represented by an $n \times n$ adjacency matrix A . Let $A_{i,j}$ ($1 \leq i, j \leq n$) be an element of matrix A .

$$A_{i,j} = \begin{cases} 1 & \text{if } i \text{ is adjacent to } j, \text{ and} \\ 0 & \text{if } i \text{ is not adjacent to } j. \end{cases} \quad (1)$$

Since $A_{i,i}$ represents a link to itself, $A_{i,i} = 0$. Let the order of node i be the Hamming weight of the i -th row (or i -th column). From the symmetry of matrix A , $A_{i,j} = A_{j,i}$

holds. The node with the highest order is defined as a “hub node”. Let λ be the eigenvalue of A , the following characteristic equation is derived.

$$\det(\lambda I - A) = 0 \quad (2)$$

Since the characteristic equation is n -dimensional, the eigenvalues can have m ($1 \leq m \leq n$) different values. Let $\lambda_{max}(A)$ be the maximum value of λ . The value of $\lambda_{max}(A)$ is a characteristic of the connection density between hub nodes. It also has the property of the “diffusion rate” of the topology map.

2.3 Laplacian Matrix

The topology map G can also be represented by a Laplacian matrix L . Let $L_{i,j}$ ($1 \leq i, j \leq n$) be an element of matrix L .

$$L_{i,j} = \begin{cases} d_i & \text{if } i = j \\ -1 & \text{if } i \text{ is adjacent to } j, \text{ and} \\ 0 & \text{if } i \text{ is not adjacent to } j, \end{cases} \quad (3)$$

where d_i denotes the order of the i -th node. The eigenvalues of L are also derived using Eq. (2), as in the adjacency matrix. Thus, there are m ($1 \leq m \leq n$) eigenvalues of L as follows.

$$0 = \lambda_1(L) \leq \lambda_2(L) \leq \dots \leq \lambda_{max}(L) \quad (4)$$

The minimum value $\lambda_1(L)$ is always equal to 0. The second minimum value $\lambda_2(L)$ indicates the algebraic connectivity of the topology map. When this value is large, the topology map has high connectivity. The maximum value $\lambda_{max}(L)$ indicates the difficulty of connectivity delays. The synchronization of the topology map can be evaluated by the ratio $R = \lambda_2(L)/\lambda_{max}(L)$.

3 Proposed Attack Strategy

3.1 Topology Map of AS

On the Internet, it is necessary to accurately grasp and maintain the destination of packets to be communicated. Various research organizations, such as RIPE [14] and the routeview project by the University of Oregon [9], have collected and published information on AS. This information is widely used for the actual maintenance of network and research activities. In this paper, we use log information from the routeview project for our computer simulations. There are two types of AS; Cisco products and Zebra products. In this paper, we focus on Zebra logs to derive a topology map of AS (Fig. 1).

```

TIME: 04/25/22 04:00:00
TYPE: TABLE_DUMP_V2/IPV4_UNICAST
FROM: aaa.bbb.ccc.ddd AS2222
ORIGIN: IGP
ASPATH: 2222 3333 4444 5555 5555 5555
NEXT_HOP: aaa.bbb.ccc.ddd
ATOMIC_AGGREGATE
AGGREGATOR: AS1234 ccc.ddd.eee.fff
COMMUNITY: 4444:222 4444:123 3333:111

```

Fig. 1. An example of AS log

We focus on ASPATH which is a list of AS numbers that the routing information has passed through. It has two important functions. First, it is used to select a route with a small number of paths when there are multiple routes to the same destination. Second, it can detect loops in a route. In these ways, a set of ASPATH shows routing information in a network. Therefore it can be used to derive a topology map among ASs. Although it contains redundant routes, these are held to derive a topology map because they have the characteristic of the network. On the other hand, duplicate routes can be omitted.

3.2 Attack Strategy

Our proposed attack strategy is defined by the combination of scenarios and tactics. In this paper, we focus on the following two scenarios.

Scenario-1 Spread of malware and disinformation

Scenario-2 Concentration and confusion of information sharing

Scenario-1 is easy to understand and a typical case of a network attack, so we omit the details. Scenario-2 is to generate the differentials in information sharing between the target area and others and make confusion among them. This scenario is also based on one of the important characteristics of Internet technology such as the immediacy of information sharing. By using this characteristic, we can generate a threshold of intentional diffusion of information. This scenario is similar to the spread of rumors, but it is different from such scenarios in the point that the difference in the spread of different information is generated deliberately. The effectiveness of these attack scenarios can be estimated by the “malicious topology map” derived from attack simulations. The effectiveness of Scenario-1 is related to the characteristic of “Spread of speed” and Scenario-2 is related to “Convergence” respectively [10, 16].

On the other hand, network attack has various tactics such as DDoS, XSS, down of services, construction of rogue servers, and so on. These tactics can affect the topology map and change its characteristics. Therefore, the adversary group can choose an attack scenario and discuss its effectiveness by selecting tactics. In this paper, we consider the following three tactics and simulate their effectiveness in the derived malicious topology map.

Tactics-1 Stopping ASs

Tactics-2 Rerouting by setting fake ASs

Tactics-3 Combination of Tactics-1 and Tactics-2

Tactics-1 can be achieved by a well-known attack such as DDoS. Tactics-2 can be achieved by setting fake AS or rewriting ASPATH. We can set any number of stopped ASs, fake ASs, and newly generated ASPATH and links. As a result, the combination of tactics and conditions makes an exponential number of kinds of malicious topology maps. However, the computer resources limit the search field, so we determine the most effective one computing with $\lambda_{max}(A)$ and R of the “initial topology map” which is before the attack in the computationally feasible range. The procedure of our proposed attack strategy is as follows.

Step-1 Collect AS logs.

Step-2 Extract AS path information from logs.

Step-3 Derive the initial topology map.

Step-4 Run Tactics-1 ~ Tactics-3 simulations for both scenarios by brute force search.

Step-5 Select the best malicious topology map in Step-4 as the best tactic for the scenario.

4 Example Execution

4.1 Preparation (Step-1 ~ Step-3)

We show an example derivation of attack strategy using Zebra log from [9]. In this example, we use “rib.20220425.0400.bz2” to extract AS path information as follows.

```
$ bgpdump “rib.20220425.0400.bz2” |grep -w “AS number” |grep -v “SEQ” |grep -v “AGG” |grep -v “COM” |tr -d “ASPATH:”
```

As a result, we have different 73,806 ASs and unique 177,109 links. Due to the limited computational power, we extract 69 nodes and 93 links among them as the initial topology map (see Fig. 2). Note that details on how we chose ASs and their links are omitted because of ethical reasons.

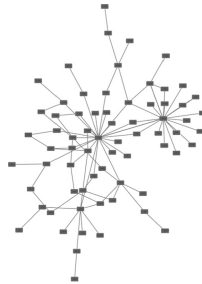


Fig. 2. Initial topology map of example attack

4.2 Simulation of Tactics and Result (Step-4)

From the results of Sect. 4.1, we can calculate $\lambda_{max}(A) = 5.3307$ and $R = 0.002127$. Due to the limited computational power (see Table 1), we execute the following total of eight tactics.

Table 1. Computer environment used in the simulation

OS	Windows 10 Home
Compiler	Python 3.6
CPU	Intel(R) Core(TM) i7-9700K CPU @ 3.60 GHz 3.60 GHz
Memory	64 GB

Tactics-1.1 Stop one AS.

Tactics-1.2 Stop two ASs.

Tactics-1.3 Stop three ASs.

Tactics-2.1 Increase one fake AS and connect it with two nodes.

Tactics-2.2 Increase one fake AS and connect it with three nodes.

Tactics-2.3 Increase two fake ASs and connect them with two nodes from each.

Tactics-3.1 Stop one AS and increase one fake AS. The increased fake AS is connected by two nodes.

Tactics-3.2 Stop two ASs and increase one fake AS. The increased fake AS is connected by two nodes.

The results of each tactics are shown in Fig. 3, Fig. 4, and Table 2. Note that detailed information on the stopped AS number and position of fake AS and links is omitted because of ethical reasons.

Table 2. Result of each tactic

Topology	$\lambda_{max}(A)$	R	Time (sec)
Initial topology map	5.3307	0.002127	0.28
Tactics-1.1	5.3307	0.003178	11.45
Tactics-1.2	5.3305	0.006895	411.72
Tactics-1.3	5.3304	0.007352	20,245.80
Tactics-2.1	5.5164	0.002127	450.44
Tactics-2.2	5.5961	0.002128	21,600.65
Tactics-2.3	5.6983	0.002127	330,594.43
Tactics-3.1	5.5164	0.003178	160,131.32
Tactics-3.2	5.5961	0.002128	307,582.62

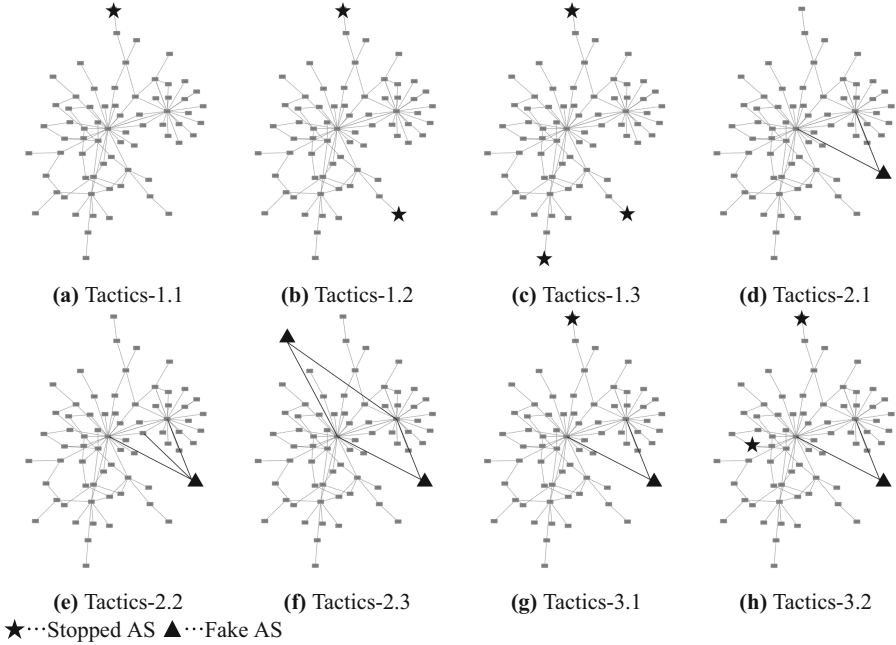


Fig. 3. Scenario-1: Spread of malware and disinformation

4.3 Select of Attack Strategy (Step-5)

4.3.1 Scenario-1

From the results of Scenario-1 (Fig. 3), we can find the following two properties of tactics. First, in Tactics-1, the eigenvalues are nearly one of the initial topology map even if the number of stopped ASs increases. Second, in Tactics-2 and Tactics-3, the eigenvalues increase as the number of generated links and the number of settings fake ASs increase. Therefore the more number of them, we can have the more effective of Scenario-1. From the above consideration, we can conclude that for Scenario-1 against the initial topology map (Fig. 2), Tactics-2.3 is the best.

The advantage of Tactics-2.3 is only setting fake ASs and the necessary technical level is quite low. The adversary group can succeed by only advertising its fake BGP without manipulating any other AS. However, it is necessary that network corporations, administrators, and are joining the adversary group. Since it is difficult to find such attack infrastructure, the attack effectiveness will be continued.

4.3.2 Scenario-2

From the results of Scenario-2 (Fig. 4), we can find the following two properties of tactics. First, in Tactics-1, the eigenvalues increase as the number of stopped ASs. Second, in Tactics-2, the eigenvalues are nearly one of the initial topology map even if the number of generated links and setting fake ASs increase. Therefore the more number of

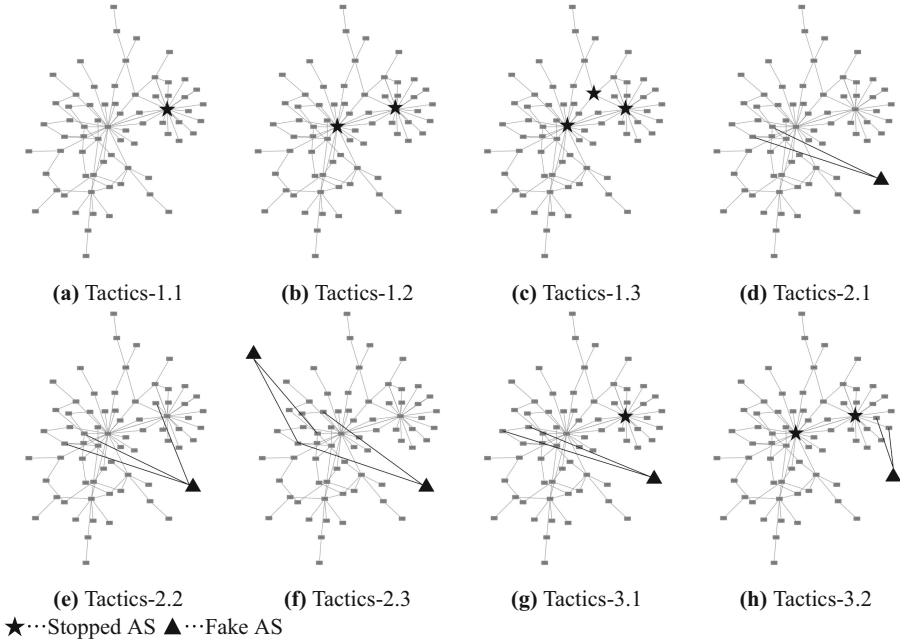


Fig. 4. Scenario-2: Concentration and confusion of information sharing

stopped ASs, we can have the more effective of Scenario-2. From the above consideration, we can conclude that Tactics-1.3 is the best tactics. Therefore it is predicted that stopping many ASs will increase the attack effectiveness.

The advantage of Tactics-1.3 is only stopping many ASs. The adversary group can also succeed by only rewriting ASPATH of the target AS. Such attacks are necessary for hostile internal collaborators or administrators, it is difficult to find such unauthorized operations and the attack effectiveness will be continued.

4.4 Consideration

4.4.1 Fake as or New Links?

Increasing new links is considered to have a big influence on changing the characteristic of the topology map. However, from the result of Scenario-1 (see Sect. 4.3.1), it turns out that this consideration may not be right. Comparing Tactics-2.2 and Tactics-2.3, it is more effective to increase the number of setting fake ASs than to increase the number of generated links. It is considered important to increase the number of fake ASs with fewer links.

4.4.2 Is Tactics-3 Effective?

Tactics-3 is very powerful but it is not chosen for both of the Scenarios. From Table 2, it is obvious that the effectiveness of Tactics-3 is not so good. However, from the view-

point of countermeasure, Tactics-3 may be effective. In particular for Scenario-2, since Tactics-3.2 holds the initial characteristic of the topology map. It is expected to be possible to return the network from the attack. This is considered to be the equivalent case of Ukraine that introduces the Starlink [18]. We can see that setting friendly ASs and rewriting ASPATH become a countermeasure. As a result, it may be appropriate to consider Tactics-3 as a defense method than an attack method.

5 Discussion

5.1 Feasibility of Proposed Attack Strategy

In general, even an attacker with standard abilities can succeed in stopping ASs and setting fake ASs in some way. However, it is unlikely that the attack will succeed, as it is necessary to be fortunate to have effective vulnerabilities and to obtain highly confidential information. On the other hand, our proposed attack strategy is based on publicly available information about ASPATH and is guaranteed to be reliable. Furthermore, since the attack is achieved simply by advertising BGP, it does not require high techniques. In this way, we can conclude that the feasibility of our proposed strategy is extremely high because it can achieve reliable results with fewer technical problems than conventional cyber attacks. However, it differs significantly from conventional cyber attacks in that it requires an internal collaborator. There are many ways to get such internal collaborators beyond cyber attacks (Intimidation, kidnapping, etc.). Problems with these are not limited to cyber security but are related to organizational management and personnel affairs. Therefore, although our proposed method is technically easy to execute, there are many topics to be resolved in the organization of adversary groups.

5.2 Effectiveness of Stopped Terminal AS

Intuitively, AS, which is a hub node, is a target with a high attack value. Similarly, since the AS of the hub node is important, it is considered that the organization that manages it is also highly implementing organizational defense measures as described in the previous section. Therefore, it is not considered a good tactic to target such AS. On the other hand, the terminal AS is less important and is likely to be the target. In fact, simulations of bank bankruptcies show that the bankruptcies of many small banks have a greater impact than the bankruptcies of megabank [7].

Based on this idea, review the attack in Scenario-2. Especially in Tactics-1.2, when the terminal AS is targeted, it is possible to easily obtain twice as much R as the initial topology map (see Fig. 5 and Table 3). It should be noted that this is a result of the shape of the topology map being concentrated in the center and does not indicate the effectiveness of the attack. As mentioned above, the terminal AS is considered to be easy to attack, but it should be noted that the intended attack result may not be obtained.

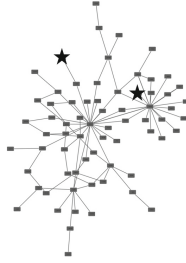


Fig. 5. Location to stop the terminal ASs

Table 3. Effectiveness of Tactics-1.2 when stop the terminal ASs

Topology	R
Initial topology map	0.002127
Tactics-1.2	0.004213

5.3 Appropriate Values of $\lambda_{max}(A)$ and R

In this paper, we judge that the case where $\lambda_{max}(A)$ and R are the most increased compared to ones of the initial topology map is effective. However, it can be the topology map with enough $\lambda_{max}(A)$ and R for the attack from the initial or even a small value obtained after applying the tactics. Therefore, it is necessary to estimate these appropriate values in advance. At the same time, it is also necessary to set these target values that can be expected to have sufficient attack effectiveness. Since the values of $\lambda_{max}(A)$ and R are determined by the shape of the topology map, these values cannot be uniquely determined. However, since the links among ASs in the world are theoretically determined by log analysis, the similarity of subspaces with a limited number of ASs can be taken into consideration. As a result, it is possible to determine the optimum value based on the similarity of topology maps. This analysis is our future work.

6 Conclusion

In this paper, we show our proposed attack strategy using AS topology map analysis and demonstrate an example attack using actual AS logs. We have shown as specific as possible, however, for ethical reasons we have not given any details of the ASPATH, such as the AS number or the location of the fake AS (the AS number to connect to). Therefore we only show our algorithms, there is one more thing we did not show concretely. It is the AS number that is the target of the attack (Scenario-1) and the point where the network is divided (Scenario-2). From the current research results, the attack target and the division point are determined from the result of topology map analysis, which may lead to undesired results in some cASs. In future work, we improve it so that we can choose the tactics after determining the target. As mentioned in Sect. 5.3, we expect that this goal can be achieved by performing a nested analysis that divides the topology map of ASs around the world and considers them as new topology maps.

References

1. arsTECHNICA. Russian-controlled telecom hijacks financial services' internet traffic (2017). <https://arstechnica.com/security/2017/04/russian-controlled-telecom-hijacks-financial-services-internet-traffic>
2. Demchak, C.C., Shavitt, Y.: China's maxim-leave no access point unexploited: the hidden story of China telecom's BGP hijacking. *Mil. Cyber Affairs* **3**(1), 7 (2018)
3. Douzet, F., Pétiinaud, L., Salamatian, L., Limonier, K., Salamatian, K., Alchus, T.: Measuring the fragmentation of the internet: the case of the border gateway protocol (BGP) during the Ukrainian crisis. In: 12th International Conference on Cyber Conflict, pp. 157–182. IEEE (2020)
4. IANA. Internet assigned numbers authority. <https://www.iana.org/>
5. Apostolaki, M., Zohar, A., Vanbever, L.: Hijacking bitcoin: routing attacks on cryptocurrencies. In: IEEE Symposium on Security and Privacy, pp. 375–392. IEEE (2017)
6. Wübbeling, M., Meier, M.: Reclaim your prefix: mitigation of prefix hijacking using IPsec tunnels. In: 42nd IEEE Conference on Local Computer Networks, pp. 330–338. IEEE (2017)
7. Zamami, R., Namatame, A.: Systemic Risk on least susceptible network. *LNEMS* (2013)
8. RIPE NCC. Youtube hijacking: a RIPE NCC RIS case study (2008). <http://www.ripe.net/internet-coordination/news/industry-developments/youtube-hijacking-a-ripe-ncc-ris-case-study>
9. University of Oregon. Route views project. <http://www.routeviews.org/routeviews/>
10. Pastor-Satorras, R., Vázquez, A., Vespignani, A.: Dynamical and correlation properties of the internet. *Phys. Rev. Lett.* **87**(25), 258701 (2001)
11. Salvador, P.: Client side localization of BGP hijack attacks with a quasi-realistic internet graph. In: Obaidat, M.S., Cabello, E. (eds.) ICETE 2017. CCIS, vol. 990, pp. 1–15. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11039-0_1
12. Sermpezis, P., et al.: ARTEMIS: neutralizing BGP hijacking within a minute. *CoRR*, abs/1801.01085 (2018)
13. Perazzo, P., Arena, A., Dini, G.: An analysis of routing attacks against IOTA cryptocurrency. In: IEEE International Conference on Blockchain, pp. 517–524. IEEE (2020)
14. RIPE. RIPE NCC. <https://www.ripe.net/>
15. Rojo, O., Soto, R.: The spectra of the adjacency matrix and Laplacian matrix for some balanced trees. *Linear Algebra Appl.* **403**, 97–117 (2005)
16. Gomez, S., Diaz-Guilera, A., Gomez-Gardenes, J., Perez-Vicente, C.J., Moreno, Y., Arenas, A.: Diffusion dynamics on multiplex networks. *CoRR*, abs/1207.2788 (2012)
17. Shapira, T., Shavitt, Y.: AP2Vec: an unsupervised approach for BGP hijacking detection. *IEEE Trans. Netw. Serv. Manag.* **1** (2022)
18. SpaceX. Starlink. <https://www.starlink.com/>
19. Tanaka, H.: Network counter-attack strategy by topology map analysis. In: Ray, I., Gaur, M.S., Conti, M., Sanghi, D., Kamakoti, V. (eds.) ICISS 2016. LNCS, vol. 10063, pp. 243–262. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49806-5_13
20. Wu, C.W.: On Rayleigh-Ritz ratios of a generalized Laplacian matrix of directed graphs. *Linear Algebra Appl.* **402**, 207–227 (2005)



Improving Classification Accuracy by Optimizing Activation Function for Convolutional Neural Network on Homomorphic Encryption

Kohei Yagyu, Ren Takeuchi, Masakatsu Nishigaki, and Tetsushi Ohki^(✉)

Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan
{yagyu,takeuchi}@sec.inf.shizuoka.ac.jp,
{nisigaki,ohki}@inf.shizuoka.ac.jp

Abstract. A secure machine learning technology that performs prediction while encrypting data using homomorphic encryption is being developed. However, Convolutional Neural Networks (CNN) on homomorphic encryption cannot use general non-linear activation functions. Thus, the classification accuracy is low. We proposed a novel method to create an activation function that improves the classification accuracy by performing a pre-training optimization on the coefficients of the polynomial approximation of the Mish function. We confirmed the improvement of classification accuracy for MNIST, Fashion-MNIST, and CIFAR-10 by optimizing the Mish function through pre-training. The classification accuracy can be improved by 4.27% for CIFAR-10. Furthermore, we showed that classification accuracy improves for Fashion-MNIST and CIFAR-10 even when different networks and datasets are optimized by pre-training the activation function. These results show that the activation function of CNNs on a homomorphic encryption can be optimized to improve classification accuracy.

1 Introduction

The use of Machine Learning as a Service (MLaaS), which provides services that enable machine learning inference in the cloud, has recently attracted much attention. Users performing inference on the cloud. It needs to share confidential data with the server, such as financial or medical information [8]. However, uploading such sensitive data to the cloud in plaintext poses a security risk since the data can be accessed from the server side.

Homomorphic Encryption (HE) preserve privacy by fully supporting homomorphic operations over encrypted data [9]. However, there are limitation to HE: it cannot perform division, comparison, and exponentiation operations. The Rectified Linear Unit(ReLU) and Sigmoid functions, which are commonly used as activation functions in general CNNs, cannot be used on HE. Therefore, Gilad's et al. [10] used a square-function as the activation function, even though it had a low classification accuracy.

Ishiyama et al. [12] created an activation function with a high classification accuracy that can be operated on HE using the least-squares method to approximate the Swish function [15] to a fourth-order polynomial. This method has a low classification accuracy associated with polynomial approximation. Therefore, there is room for further study on activation function approximation with a higher classification accuracy.

In this paper, we propose a method to generate an activation function based on a polynomial approximation and accuracy optimization to maximize classification accuracy for a prepared dataset. We will also investigate methods to generate activation functions that further improve classification accuracy without increasing the order of activation functions. To evaluate our model, we performed pre-training optimization on a polynomial activation function and evaluated its classification accuracy on MNIST [19], Fashion-MNIST [18], CIFAR-10 [1], and CIFAR-100 [1] which are widely used datasets for benchmarking. The results show that the classification accuracy is higher than that of the activation function approximated polynomially by the least-squares method. The classification accuracy of Fashion-MNIST and CIFAR-10 is further improved even when the network and the dataset to be optimized are different from the ones used in training.

The rest of the paper is organized as follows. We describe our research and related research questions in Sect. 2. In Sect. 3, we introduce our method of improving classification accuracy. We discuss the classification accuracy results with our activation functions in Sects. 4. Finally, we give the conclusions of our work and some recommendations for future work in Sect. 5.

2 Related Work

Gilad-Bachrach et al. [10] implemented the world’s first CNN on HE. They showed that it is possible to obtain inference results while preserving privacy by encrypting the model parameters and the images to be inferred using HE. A square-function was used as the activation function, Although It achieves 99% classification accuracy for MNIST, its output changes significantly depending on the value of the input, resulting in a low classification accuracy.

Hesamifard et al. [11] performed learning and inference in CNNs on HE by performing polynomial approximations of the activation function. They used Chebyshev polynomials and Taylor expansions as methods for polynomial approximation. The ReLU function was approximated by Chebyshev polynomial, but the classification accuracy was low. In contrast, approximating the ReLU function with a Taylor expansion resulted in better classification accuracy. Similar approximations to the Sigmoid and Tangent Hyperbolic Function, however, resulted in a lower classification accuracy. The datasets used to evaluate the classification accuracy were MNIST and CIFAR-10.

Chabanne et al. [6] performed polynomial approximations of the ReLU function at various degrees by the least-squares method to verify the classification accuracy with deeper layers of CNNs on HE. Using Average Pooling and Batch

Normalization layers for the CNN, they achieved 99.30% classification accuracy on MNIST. However, the shallow layer model yielded a much lower classification accuracy. The method used in this study was computationally expensive, as more layers were needed to achieve higher classification accuracy.

Ishiyama et al. [12] performed a polynomial approximation, using the least-squares method, of the Swish function, which was versatile and had a high classification accuracy. It achieved a classification accuracy of 99.29% for MNIST and 80.47% for CIFAR-10, which is a higher classification accuracy than the other studies reported here. The classification accuracy of the Swish function and the approximate Swish function differed by approximately 2%, indicating that there is room for improvement in approximation methods as there is a gap in classification accuracy between before and after approximations.

All the above studies used a polynomial that approximates the square-function and an activation function commonly used in CNNs as the activation function. Chebyshev polynomial, Taylor expansion, and the least-squares methods are all polynomial approximations of the activation function, so low-degree polynomials are insufficient to approximate the original activation function, resulting in a large gap in classification accuracy. Thus, we focused on methods to generating activation functions based on the actual classification accuracy.

In this paper, we show that optimizing the coefficients of the polynomial approximations of activation functions by pre-training leads to a better classification accuracy on benchmarking datasets.

3 Proposed Method

The proposed method in this paper can be divided into two main steps.

1. The activation function and the Mish function, which have higher classification accuracy than the Swish function, were approximated polynomially by the least-squares method.
2. Hyperparameter tuning was performed to maximize the classification accuracy on the prepared dataset. Specifically, Optuna’s Tree-structured Parzen Estimator (TPE) [3] maximizes classification accuracy by searching for each coefficient of the Mish function generated by a polynomial approximation.

First, in the polynomial approximation of the Mish function in step 1., a polynomial approximation was performed by the least-squares method similar to Ishiyama et al. [12]. We decided to use the Mish function [14] as the activation function for polynomial approximation, which is considered more versatile and has a higher classification accuracy than ReLU and Swish function.

Next, in the process of hyperparameter tuning by TPE for the coefficients of the activation function in step 2., the coefficients of the polynomial activation function were searched for based on the classification accuracy of the actual dataset output by a pre-training process. In practice, TPE optimization was performed on the approximated Mish function. To further improve the accuracy of

the approximation to the Mish function, optimization is performed by narrowing the search range for the polynomial coefficients. In this way, we can expect the activation function to have high classification accuracy, similar to that of the Mish function. Furthermore, optimizations with a narrower search range are expected to allow generating commonly used activation functions such as ReLU and Sigmoid functions due to the reduced dependence of the optimized Mish function on the network or the dataset on which the activation function was optimized.

3.1 Polynomial Approximation of the Mish Function

In this study, a polynomial approximation of the Mish function was performed by the least-squares method which had the highest classification accuracy in the other studies [11, 12], was utilized as the approximation method. In the study by Ishiyama et al. [12], the degree of a polynomial was set to 4, which had the highest classification accuracy, and the range of inputs to be approximated was set to $[-6, 6]$. The degree and the range of inputs to be approximated are parameters for the approximation of the Swish function, but due to the high similarity between the Swish and Mish functions, these apply to the Mish function as well. We base our expression construction on the polynomial equation in [12]. The approximate Mish function is shown in equation (1) and the graph is shown in Fig. 1.

$$f(x) = -0.0021023x^4 + 0.14866987x^2 + 0.51008878x + 0.16863171 \quad (1)$$

3.2 Accuracy Optimization

We used TPE optimization to find the coefficients of the approximate Mish function. Specifically, the coefficients of the activation function in the network to be optimized were modified and trained for a given number of epochs using the training data. Then, the coefficients with the highest classification accuracy for the data not used for training were searched. The search ranges for the coefficients were set to $[-1, 1]$, $[-0.1, 0.1]$, $[-0.01, 0.01]$, and $[-0.001, 0.001]$. The datasets used for optimization by TPE were MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100. For MNIST and Fashion-MNIST, 48,000 data were used for training, and 12,000 data were used to evaluate the classification accuracy during optimization. The remaining 10,000 data were used to validate the classification accuracy of the activation function after optimization to avoid overfitting. For CIFAR-10 and CIFAR-100, 40,000 images were used for training, and 10,000 images were used to evaluate the classification accuracy during optimization. The number of training sessions was set to 100, and the number of epochs was set to 10 to account for the training cost. Batch size was set to 128, learning rate to 0.01, and optimizer to AdaDelta.

The respective network architectures for MNIST (MNIST), Fashion-MNIST (Fashion-MNIST), and CIFAR-10 (CIFAR-10) used in the optimization are

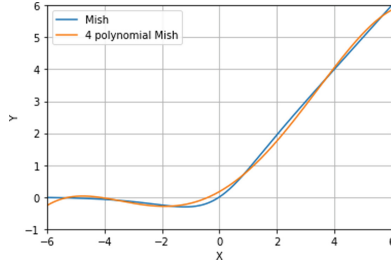


Fig. 1. Mish function approximated by a 4th degree polynomial

shown in Table 1 and Table 2. The network used for the optimization of CIFAR-100 (CIFAR-100) was based on VGG16 [16] with the activation function replaced by a polynomial activation function. The idea is to verify if the network or the dataset to be optimized is different from the one used in training. We verified that it is commonly used for benchmarking purposes and that the optimized activation function can produce a general-purpose activation function without considering the depth-differences of the layers. Since there are limitations on verification networks and HE, it is expected that simple networks will generate more versatile activation functions than complex networks that cannot be operated on HE.

The equations of the optimized Mish function created for MNIST are shown in Table 3, and the graphs of its polynomial activation functions are shown in Fig. 2 (a). The equations of the optimized Mish function created for MNIST using Fashion-MNIST are shown in Table 4, and the graphs of its polynomial activation functions are shown in Fig. 2 (b). The equations of the optimized Mish function created for CIFAR-10 are shown in Table 5, and the graphs of its polynomial activation functions are shown in Fig. 2 (c). The equations of the optimized Mish function created in VGG16 network using CIFAR-100 dataset are shown in Table 6, and the graphs of its polynomial activation functions in Fig. 2 (d). From Fig. 2, it is clear that, by narrowing the coefficient search range, the approximated Mish function is more similar to Mish function.

Table 1. Network layers MNIST and Fashion-MNIST

Layer	Parameters	Output size
Convolution	5 filters of size 5×5 , (2,2) stride	$12 \times 12 \times 5$
Batch normalization		$12 \times 12 \times 5$
Activation function		$12 \times 12 \times 5$
Convolution	50 filters of size 5×5 , (2,2) stride	$4 \times 4 \times 50$
Batch normalization		$4 \times 4 \times 50$
Activation function		$4 \times 4 \times 50$
Fully connected	10 units	$1 \times 1 \times 10$

Table 2. Network layer for CIFAR-10

Layer	Parameters	Output size
Convolution	40 filters of size 5×5 , (2,2) stride	$16 \times 16 \times 40$
Batch normalization		$16 \times 16 \times 40$
Activation function		$16 \times 16 \times 40$
Average Pooling	Pool size 5×5 , (2,2) stride	$6 \times 6 \times 40$
Convolution	80 filters of size 3×3 , (1,1) stride	$6 \times 6 \times 80$
Batch normalization		$6 \times 6 \times 80$
Activation function		$6 \times 6 \times 80$
Fully connected	10 units	$1 \times 1 \times 10$

Table 3. Equations of optimized Mish function created for MNIST

Range	Activation function
$[-1, 1]$	$f(x) = -0.01442524x^4 + 0.00158332x^2 + 0.88826944x - 0.01275787$
$[-0.1, 0.1]$	$f(x) = -0.01069226x^4 + 0.11951754x^2 + 0.56319683x + 0.09321102$
$[-0.01, 0.01]$	$f(x) = -0.00831687x^4 + 0.14990016x^2 + 0.50010013x + 0.16206926$
$[-0.001, 0.001]$	$f(x) = -0.00201326x^4 + 0.14901227x^2 + 0.50967517x + 0.1694176$

Table 4. Equations of optimized Mish function created for Fashion-MNIST

Range	Activation function
$[-1, 1]$	$f(x) = 0.01809696x^4 - 0.59336015x^2 + 1.17978706x + 0.44916505$
$[-0.1, 0.1]$	$f(x) = -0.00735186x^4 + 0.19722884x^2 + 0.45351870x + 0.07937917$
$[-0.01, 0.01]$	$f(x) = 0.00033206x^4 + 0.14147889x^2 + 0.50264848x + 0.16690980$
$[-0.001, 0.001]$	$f(x) = -0.00162847x^4 + 0.14815714x^2 + 0.51101085x + 0.16805804$

Table 5. Equations of optimized Mish function created for CIFAR-10

Range	Activation function
$[-1, 1]$	$f(x) = -0.01410303x^4 + 1.06744797x^2 + 0.62838373x + 0.31748578$
$[-0.1, 0.1]$	$f(x) = -0.00252074x^4 + 0.22103484x^2 + 0.45203568x + 0.15522165$
$[-0.01, 0.01]$	$f(x) = -0.00147005x^4 + 0.14283672x^2 + 0.50519885x + 0.17068954$
$[-0.001, 0.001]$	$f(x) = -0.00115840x^4 + 0.14767311x^2 + 0.51018185x + 0.16885427$

4 Evaluation

We evaluated our novel optimization-based activation function to show its effectiveness. In Sect. 4.1, we explain computing environment, implementations, and training condition of the activation functions. In Sects. 4.2 and 4.3, we report our evaluation results for classification accuracy and transferability of the proposed activation function.

Table 6. Using CIFAR-100 on the VGG16, the optimized Mish function equations

Range	Activation function
$[-1, 1]$	$f(x) = -0.65204331x^4 + 1.00256451x^2 + 1.40638317x + 0.65899197$
$[-0.1, 0.1]$	$f(x) = -0.06910764x^4 + 0.11355908x^2 + 0.54662747x + 0.15644139$
$[-0.01, 0.01]$	$f(x) = -0.00069412x^4 + 0.14754837x^2 + 0.50595350x + 0.16594631$
$[-0.001, 0.001]$	$f(x) = -0.00134593x^4 + 0.14913091x^2 + 0.51026693x + 0.16789240$

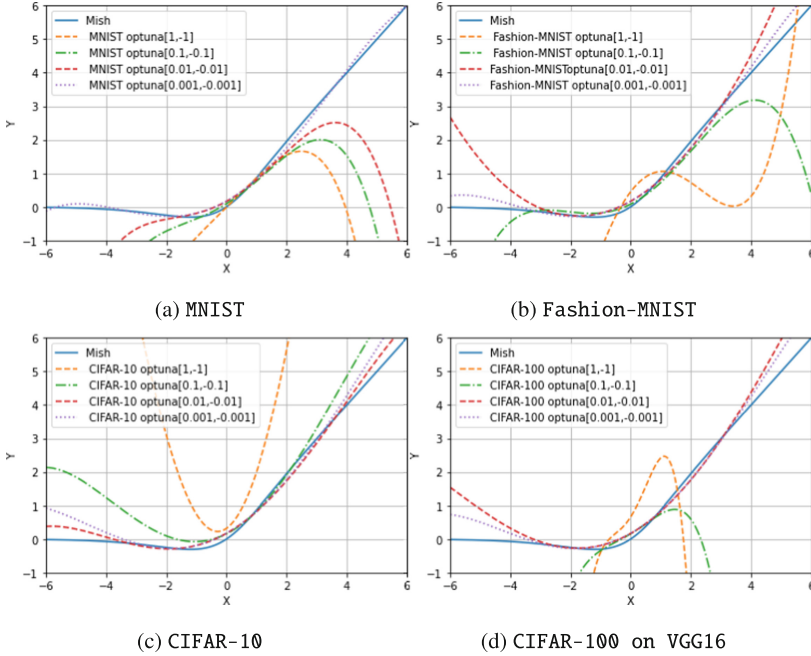


Fig. 2. Graphs of polynomial activation functions

4.1 Evaluation Setup

Computing Environment. We used a RyzenTM ThreadripperTM 3990X with 64 cores and 128 GB RAM for inference.

Dataset. As described in 3.2, we used MNIST dataset, Fashion-MNIST dataset and CIFAR-10 datasets for evaluation. For each dataset, we split 10,000 images into test images and the rest into training images.

Implementation. The implementation of the model was done using Tensorflow [2], and the library to perform inference on the HE was designed with a graph compiler, nGraph-HE [5], as the back-end. Using nGraph-HE, it is possible to easily implement CNN on HE, and the graph compiler speeds up the execution. We used SEAL [13] for Cheon-Kim-Kim-Song (CKKS) scheme [7], which was implemented under the nGraph-HE environment. We used packing [17] for evaluating the HE and SIMD for inference, which significantly improved the execution speed. Additionally, as the number of multiplications has a large impact on computation time and memory utilization in the HE processes, we adopted two methods to reduce the number of multiplications: fusing convolutional layer with Batch Normalization and reducing the coefficients of the polynomial activation function [4, 5]. As the library of HE, SEAL, is not GPU-compatible, we

performed the evaluation on CPU. To increase the execution speed, we used 64-core OpenMP parallel processing. The encryption parameters: N was 8192, Scale-Factor was 24, q/bits was 180 for MNIST and Fashion-MNIST, and 204 for CIFAR-10.

Training. In our experiments, MNIST, Fashion-MNIST, and CIFAR-10 were trained on plaintext. The number of training epochs was set to 100, and the other parameters were the same as described in Sect. 3.2. We applied Data Augmentation during training. The activation functions used for each network were square-function, Mish function, Swish function, approximated Mish function, approximated Swish function [12], and proposed Mish function optimized by TPE. For optimizing Mish function, using accuracy optimization, we used the training dataset corresponding to the network (for example, we conducted accuracy optimization for MNIST network using MNIST dataset). We varied the search range in $[1, -1]$, $[0.1, -0.1]$, $[0.01, -0.01]$, and $[0.001, -0.001]$.

In addition, we optimized the Mish function using CIFAR-100 dataset on VGG16 for evaluating the transferability of optimized function. These cases are denoted as VGG16_Optuna, VGG16_Optuna $[1,-1]$, VGG16_Optuna $[0.1,-0.1]$, VGG16_Optuna $[0.01, 0.01]$, and VGG16_Optuna $[0.001,-0.001]$ for the corresponding search ranges. We also evaluated square-function [10] and approximated Swish function [12] for comparison, which were used to evaluate the relative accuracy of approximation and classification by the proposed method because Swish and Mish functions can only perform inference on plaintext.

Inference. We evaluated the classification accuracy on the plaintext and compared it with the accuracy on ciphertext. For MNIST and Fashion-MNIST, in both evaluations, 10,000 images were used for inference. For CIFAR-10, the number of images inferred was 1000 due to insufficient memory capacity on the computer used for the evaluation on ciphertext.

4.2 Experimental Results for Classification Accuracy

We evaluated how our activation function preserves classification accuracy. Table 7 shows the classification accuracy of the models trained with each activation function on the plaintext and ciphertext. There is a slight difference in the classification accuracy between the plaintext and the ciphertext. This is considered to be an error caused by the encryption method that approximates the real numbers in the CKKS scheme. We observed that the classification accuracy of the Mish function is higher than that of the Swish function on plaintext, and similarly, the approximated Mish function was higher classification accuracy than that of the approximated Swish function [12] on plaintext and ciphertext. We found that the optimized Mish function had a higher classification accuracy than the approximated Mish function when the search range was $[-1, 1]$. The TPE-optimized activation function showed higher classification accuracy when the search range was enlarged. The classification accuracy with the optimized

Table 7. Results for classification accuracy

Activation function	Classification accuracy on plaintext [%]			Classification accuracy on ciphertext [%]		
	MNIST	Fashion-MNIST	CIFAR-10	MNIST	Fashion-MNIST	CIFAR-10
Square-function	97.97	84.01	64.32	97.97	84.01	63.55
Mish	99.45	85.46	73.74	N/A	N/A	N/A
Swish	99.38	85.24	72.89	N/A	N/A	N/A
Approximated Mish	99.29	84.69	66.55	99.30	84.68	67.20
Approximated Swish [12]	99.13	84.46	66.12	99.11	84.44	66.50
Optuna[-1,1]	99.35	85.20	70.82	99.35	85.20	71.05
Optuna[-0.1, 0.1]	99.18	84.74	69.01	99.18	84.75	69.00
Optuna[-0.01, 0.01]	99.11	84.69	67.78	99.12	84.68	68.15
Optuna[-0.001, 0.001]	99.24	85.10	70.23	99.22	85.10	70.15
VGG16.Optuna[-1, 1]	97.10	82.38	48.83	97.07	82.38	50.37
VGG16.Optuna[-0.1, 0.1]	98.07	82.15	58.67	98.07	82.14	60.40
VGG16.Optuna[-0.01, 0.01]	98.86	84.04	67.34	98.86	84.04	67.71
VGG16.Optuna[-0.001, 0.001]	99.20	85.18	70.21	99.18	85.17	70.04

Mish function is 4.27% higher than that of the approximated Swish function for CIFAR-10.

4.3 Experimental Results for Different Networks

We evaluated the classification accuracy of the optimized activation function on different networks. Table 7 shows that activation function optimized on VGG16 using CIFAR-100 had a lower classification accuracy than the approximated Mish function for all search ranges on MNIST. However, the Fashion-MNIST and CIFAR-10 results showed that the activation function optimized on VGG16 using CIFAR-100 had higher classification accuracy than the approximated Mish function when the search range was set to $[-0.001, 0.001]$.

5 Disucussion

5.1 Effective Optimization Method for Activation Functions in CNN

From Table 7, we see that the approximate Mish function has a higher classification accuracy than the approximated Swish function in the study by Ishiyama et al. [12]. This may be due to the fact that the Mish function has better classification accuracy than the Swish function. The optimized Mish function outperformed the approximated Mish function in classification accuracy only for the search range $[-1, 1]$. This indicates that narrowing the search range does not necessarily increase the classification accuracy when optimizing using TPE. Furthermore, when the target network or dataset is clear, optimizing with a wide search range and overtraining will result in higher classification accuracy. However, as the search range $[-0.001, 0.001]$ is superior to the search range $[-0.1,$

0.1] and $[-0.01, 0.01]$ in classification accuracy, it can be said that classification accuracy is sometimes improved when the search range is narrower.

The results of the evaluation by Fashion-MNIST, shown in Table 7, show that among the optimized Mish functions, the classification accuracy is improved in all cases except for the search range $[-0.01, 0.01]$. Additionally, all activation functions optimized by TPE outperform the classification accuracy of the approximate Mish function in the evaluation results by CIFAR-10. These results strongly indicate the effectiveness of the optimization in improving the classification accuracy. Meanwhile, the lack of significant improvement in classification accuracy on MNIST can be attributed to the fact that MNIST itself is an easy image classification task, and thus a network with very high classification accuracy can be constructed without considering the approximation accuracy of the activation function.

We conclude that for MNIST, Fashion-MNIST, and CIFAR-10, when the network and dataset are the same during optimization and training, a wider search range can greatly improve the classification accuracy. If the search range is set to $[-0.01, 0.01]$ or $[-0.1, 0.1]$, a narrower search range will improve the classification accuracy.

5.2 Application of Optimization to Different Networks

The classification accuracy for MNIST, presented in Table 7, shows that the optimized Mish function created on VGG16 using CIFAR-100 had a lower classification accuracy than the approximate Mish function in all search ranges. This indicates that, in MNIST, the optimized Mish functions generated on a network or dataset different from the one used in training did not improve classification accuracy. However, the classification accuracy Table 7 for Fashion-MNIST and CIFAR-10 show different results than that for MNIST. The classification accuracy of the Mish function is better than that of the approximate Mish function in Fashion-MNIST and CIFAR-10 only when the search range is $[-0.001, 0.001]$. This suggests that by narrowing the search range, even for an activation function optimized on a network or dataset different from the one used in training may improve the classification accuracy.

These results suggest that optimizing the activation function using a narrow search range can improve classification accuracy when the network and the dataset to be optimized are different from the ones used in training. For a simple identification task such as MNIST, it is difficult to improve classification accuracy by optimizing under any of the conditions.

5.3 Limitations

One limitation of this study is the large learning cost overhead of optimizing the activation function. Compared to the creation of activation functions using the least-squares method, optimization using TPE is more time-consuming. The second limitation is the verification when the degree of the activation function to

be optimized is small, which can be evaluated by generating an activation function that requires fewer multiplications and has higher classification accuracy. As we only tested the optimized activation function on VGG16 with CIFAR-100 models, we also need to perform more general validation on different datasets and networks.

To improve classification accuracy, GridSearch optimization can be considered instead of TPE optimization. However, compared to TPE, GridSearch has a time-consuming optimization problem.

6 Conclusion

In this paper, we proposed a novel method to create an activation function that improves the classification accuracy by performing a pre-training optimization on the coefficients of the polynomial approximation of the Mish function. Furthermore, when optimization of the activation function is performed on the data used to train the network, the search range for optimization was extended, which greatly contributed to the improvement of classification accuracy. The slight difference in classification accuracy between plaintext and ciphertext indicates that the CIFAR-100 encrypted with the CKKS scheme introduces errors. Fashion-MNIST and CIFAR-10 further improved classification accuracy by significantly reducing the search range during optimization, even when a different network or dataset was used for optimization than during training.

For HE, the improved classification accuracy brings the CNN closer to practical applications, where inference is performed while images are kept confidential.

Future work includes verifying the polynomial activation function when the degree of the activation function is small, verifying the optimized activation function on different datasets and networks, and verifying the optimization by GridSearch instead of Optuna's TPE optimization as a method to further improve classification accuracy.

References

1. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10 and cifar-100 datasets (canadian institute for advanced research). <http://www.cs.toronto.edu/~kriz/cifar.html>. Accessed 23 Jan 2022
2. Abadi, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 2016), pp. 265–283 (2016)
3. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631 (2019)
4. Boemer, F., Costache, A., Cammarota, R., Wierzynski, C.: ngraph-he2: a high-throughput framework for neural network inference on encrypted data. In: Proceedings of the 7th ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography, pp. 45–56 (2019)

5. Boemer, F., Lao, Y., Cammarota, R., Wierzynski, C.: ngraph-he: a graph compiler for deep learning on homomorphically encrypted data. In: Proceedings of the 16th ACM International Conference on Computing Frontiers, pp. 3–13 (2019)
6. Chabanne, H., de Wargny, A., Milgram, J., Morel, C., Prouff, E.: Privacy-preserving classification on deep neural network. *IACR Cryptol. ePrint Arch.* **2017**, 35 (2017)
7. Cheon, J.H., Kim, A., Kim, M., Song, Y.: Homomorphic encryption for arithmetic of approximate numbers. In: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017. LNCS, vol. 10624, pp. 409–437. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70694-8_15
8. Erickson, B.J., Korfiatis, P., Akkus, Z., Kline, T.L.: Machine learning for medical imaging. *Radiographics* **37**(2), 505–515 (2017)
9. Gentry, C.: A fully homomorphic encryption scheme. Stanford university (2009)
10. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning, pp. 201–210. PMLR (2016)
11. Hesamifard, E., Takabi, H., Ghasemi, M.: Deep neural networks classification over encrypted data, pp. 97–108. Association for Computing Machinery (2019). <https://doi.org/10.1145/3292006.3300044>
12. Ishiyama, T., Suzuki, T., Yamana, H.: Highly accurate CNN inference using approximate activation functions over homomorphic encryption. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 3989–3995. IEEE (2020)
13. Laine, K.: Microsoft seal. <https://github.com/microsoft/SEAL>. Accessed 30 Jan 2022
14. Misra, D.: Mish: a self regularized non-monotonic neural activation function. arXiv preprint [arXiv:1908.08681](https://arxiv.org/abs/1908.08681) 4, 2 (2019)
15. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018, Workshop Track Proceedings. OpenReview.net (2018). <https://openreview.net/forum?id=Hkuq2EkPf>
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015). <http://arxiv.org/abs/1409.1556>
17. Smart, N.P., Vercauteren, F.: Fully homomorphic simd operations. *Des. Codes Cryptogr.* **71**(1), 57–81 (2014)
18. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint [arXiv:1708.07747](https://arxiv.org/abs/1708.07747) (2017)
19. LeCun, C.Y.: Cortes: mnist handwritten digit database. <http://yann.lecun.com/exdb/mnist/>. Accessed 23 Jan 2022



An Attention Mechanism for Visualizing Word Weights in Source Code of PowerShell Samples: Experimental Results and Analysis

Yuki Mezawa^(✉) and Mamoru Mimura^(ID)

National Defense Academy, Yokosuka, Japan
{em60014,mim}@nda.ac.jp

Abstract. Methods that utilize AI as a detection technique for malware have been studied, and this is also true for the detection of malicious PowerShell scripts. Previous studies have proposed models that use deep learning and machine learning to detect malicious PowerShell scripts and have achieved high detection rates. However, these studies have focused on improving the detection rate of malicious PowerShell scripts. Therefore, the reasons why the detection models are determining malicious and benign PowerShell samples are unclear. In this study, we use the attention mechanism to visualize the words that are important to the malicious PowerShell scripts detection model. Then, we analyze the distribution of important words for each sample classification result. The experimental results show that there were significant differences in the words that classify benign or malicious PowerShell scripts. In addition, the misclassified samples often contain words that were emphasized in the opposite class.

1 Introduction

Wide varieties of malware are created and distributed around the world on a daily basis. Malicious PowerShell script is becoming increasingly important among the many types of malware. It has been reported that malicious PowerShell scripts accounts for 99.6% of all scripts detected in the first quarter of 2022 [9]. PowerShell is a regular tool used for Windows system administration and it has powerful functions. However, there are aspects of its powerful functions that are being abused by attackers.

In response to attacks using such malicious PowerShell scripts, previous research has proposed a method that combines natural language processing techniques and convolutional neural networks [4]. A method based on static analysis combining natural language processing techniques and machine learning models has also been proposed [7,8]. Such machine learning models and detection models using deep learning are also expected to detect unknown malicious PowerShell scripts. However, most previous studies [1,4,7,8] are aimed at improving

the accuracy of malicious PowerShell scripts detection. Therefore, it is unclear which features of PowerShell scripts are important for classification. Analyzing PowerShell scripts features that are important to the detection model across the whole dataset would contribute to improving detection rates. Therefore, this study performs a word-level static analysis of malicious PowerShell scripts using Attention mechanism on the dataset used in the previous studies. Attention mechanism is a type of deep learning. Attention mechanism allow us to visualize the weight of each token in the input data, called “attention weight”. When using attention mechanism for the static analysis of malicious PowerShell scripts, the tokens are each word in the source code. The attention weight of each words were aggregated and compared for each classification result in the dataset to analyze which words are important for classification. To the best of our knowledge, this is the first analysis of words contribution in a malicious PowerShell scripts detection model.

This paper provides the following contributions:

1. We analyzed the distribution of words that the malicious PowerShell scripts detection model emphasizes for classification by using the attention mechanism.
2. We confirmed that the words that indicate whether a PowerShell script is benign or malicious are distinctly different.
3. We found that the misclassified samples often contain words that are emphasized in the opposite class.

The structure of this paper is shown below. Section 2 introduces related work and Sect. 3 introduces related techniques. Section 4 describes the experimental method of this study, and Sect. 5 describes the experiments and results. Section 6 provides a discussion. Finally, we conclude this paper.

2 Related Work

Previous studies of AI-based malicious PowerShell scripts detection have been based on dynamic analysis and static analysis. An example of dynamic analysis is the study by Hendler et al. [4]. This study used natural language processing techniques and deep learning to detect malicious PowerShell commands. They tested 9-CNN, 4-CNN, and LSTM as deep learning architectures, and bag-of-words and 3-gram as natural language processing techniques. Then, they combined 4-CNN and 3-gram, which had the best detection rate among them, and proposed it as a Deep/Traditional Models Ensemble. An example of static analysis is the study by Tajiri et al. [7,8]. This is a method to detect malicious PowerShell scripts by analyzing PowerShell source code using natural language processing techniques and machine learning models. They used Support Vector Machine, XGBoost, and Random Forests as machine learning models, and Bag-of-Words, Latent Semantic Indexing (LSI), and Doc2Vec as natural language processing techniques. Their experiments showed that the detection model combining LSI

and XGBoost recorded the highest detection rate of 0.98. Their study also deobfuscated the PowerShell source code, and transformed the complex code to make it easier to input into the machine learning model. In addition, Choi proposed a filtering method using an attention mechanism as a countermeasure against malicious PowerShell script adversarial attacks using a Generative Adversarial Network (GAN) [1]. This study shows that even when the conventional deep learning-based detection model has a detection rate of 0% due to an adversarial attack, the detection rate can be improved to 96.5% by restoring the original malicious PowerShell using the attention mechanism. They also showed that the proposed method can also improve the detection rate of ordinary malicious PowerShell scripts.

Thus, the detection model using AI and the attention mechanism achieves a high detection rate in detecting malicious PowerShell scripts. However, it is difficult to say that enough analysis has been done to determine which parts of PowerShell script the detection model focuses on. Analyzing the factors that determine whether a sample is a benign or malicious PowerShell script by AI would contribute to improving the detection rate. Therefore, in this study, we visualize the importance of features in the sample using the Attention mechanism and analyze their distribution.

3 Related Technique

3.1 Long Short-Term Memory

Long short-term memory (LSTM) is a type of Recurrent Neural Network (RNN). The prototype of LSTM was proposed by Hochreiter et al. in 1997 [5]. It was designed to solve the vanishing gradient problem and the exploding gradient problem that RNN had. Nowadays, it has been improved and applied to speech recognition [2], handwriting recognition [3], and so on. Unlike ordinary RNN, LSTM have three gates (input gate, forget gate, and output gate) in addition to cell that is the memory part. The input gate controls the information to be newly stored in the cell. The forget gate controls the range of memory that the cell retains. The output gate controls the level to which the cell's values are used to compute outputs. The model used in this study is called a bidirectional LSTM. Normal LSTM learns from the beginning of a sentence, which is the oldest chronological order and predicts the meaning of words. However, source codes have complex word combinations. Therefore, we employed a bidirectional LSTM that can predict word meanings from the context before and after. In this study, the Keras library is used to implement LSTM.

3.2 Attention Mechanism

The attention mechanism is a type of deep learning originally developed for machine translation. The attention mechanism provides the prediction model to where “attention” should be paid to the input data. In addition, by visualizing

this attention, it is possible to disclose which parts of the input data were paid attention to and the reason for the prediction. In this study, we use what is called Self-Attention among the attention mechanisms. This is a model in which the stored and input data are the same, and learning is performed on the structure of the sentence. In addition, there is a model called Source Target-Attention. This model is mainly used for translation tasks.

4 Experimental Method

4.1 Outline

The experimental method consists of a Training Phase and a Test Phase. The data set includes training and test data, each containing samples of benign and malicious PowerShell scripts. The training data is used to train the experimental model, which is the classifier. The test data is the data used to actually classify the experimental model and measure the attention weights after classification. The validation procedure is shown in Fig. 1. First, we perform deobfuscation, data cleansing, and separation as a preprocessing step on each data set. Second, we train the experimental model through a training phase to make it classifiable. Third, we execute the Test Phase. In the Test Phase, we let the experimental model classify the test data and extract the attention weights for each word in the sample. The details of each procedure are described below.

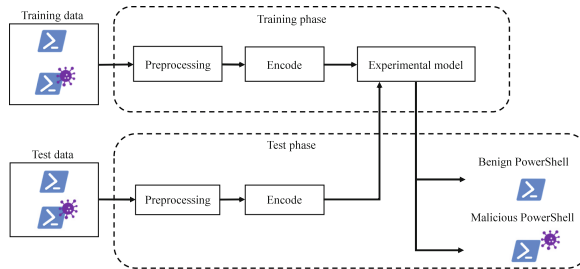


Fig. 1. Experimental method

4.2 Preprocessing

In the preprocessing step, we perform deobfuscation, data cleansing, and separations for each data set. In this process, we use regular expressions. In deobfuscation, the obfuscated parts are extracted by matching, and then unified to lowercase letters, replaced with Base64 encoding, line breaks at the end character, and so on. In data cleansing, we replace comment-outs, URLs, IP addresses, and multibyte characters, respectively. Then, we make them available as one of the features. In separations, we split strings at command and variable boundary symbols (whitespace, terminators, parentheses, and operators).

4.3 Training Phase

In the training phase, we train the experimental model that will serve as the classifier. First, the preprocessed training data is encoded with a tokenizer. This allows each word to be represented as a unique ID and the source code as a vector. Note that special tokens are inserted at the beginning and end of the sample due to the effect of the tokenizer. The encoded training data is used to train the experimental model. Trained experimental model will be able to classify samples.

4.4 Test Phase

In the Test Phase, we input the test data to the experimental model and have it classified. First, we encode the preprocessed test data with Tokenizer. In this process, unknown words may be encountered. Such words are encoded as “UNK” and recognized as specific words. Second, we let the encoded test data be classified into the experimental model. Third, we compare the attention weights of the words in each classified sample and extract the top 10 words with the highest values. Finally, the extracted words are counted to identify the most frequent words contributing to the classification.

5 Experimental Model and Evaluation Results

5.1 Experimental Model

The experimental model used in this study is a classifier that combines LSTM with an Attention mechanism. The structure of the experimental model is shown in Fig. 2. Input data is first passed to the Embedding layer, which transforms words into vector space. Since simple one-hot encoding results in a huge amount of features, Embedding is used to reduce the size. The Embedding layer passes the output 3D tensor to the Self-Attention layer. The Self-Attention layer output, with each word given an attention weight, is passed to the LSTM layer. Since the LSTM output is a 3D tensor, it is converted to a 2D tensor in the Global Max Pooling layer. The output of the Global Max Pooling layer is input to the Dense layer. The relu function is used here. The output of the Dense layer is input to the Dropout layer. The Dropout layer is effective in preventing overlearning by randomly disabling nodes. In the last layer, the labels are classified using the sigmoid function. Note that the length of the input data was set to 256, the batch size to 8, the number of epochs to 16, and the dropout rate to 0.5. Experiments were conducted with other parameters, but we did not obtain good classification results.

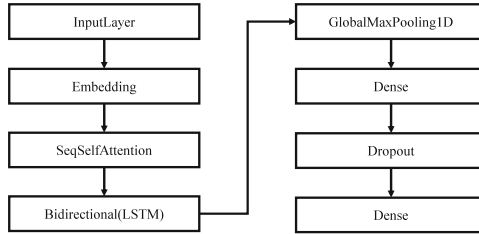


Fig. 2. Experimental model

5.2 Dataset

In this study, we were able to obtain cooperation from Tajiri et al. and used the same dataset as in their study [8]. The details of the datasets used are shown in Table 1. The dataset was created from PowerShell scripts collected from HybridAnalysis (589 scripts), PowerShell scripts collected from AnyRun (355 scripts), and benign PowerShell scripts obtained from github (5000 scripts). The collection period is between January 2019 and March 2020. All samples are publicly available on these sites. Thus, the samples were collected over a long period of time from multiple sources, and the samples are comprehensive. We used multiple security vendors for labeling the collected samples. The security vendors we used were Kaspersky, McAfee, Microsoft, Symantec, and TrendMicro. Samples that were determined to be a threat by two or more of these vendors were assigned the label malicious. Samples for which zero vendors were identified as malicious were assigned the label benign. Samples that did not fall into either category were excluded. The collected samples were divided into time series. The reason is that AI cannot be trained on unknown samples when actual malicious PowerShell scripts detection is assumed. Samples obtained from HybridAnalysis and AnyRun were split before June 2019 and after July 2019 based on time stamps. Samples obtained before June were used for training data as known samples and those obtained after July were used for test data as unknown samples. Samples obtained from github were randomly split into two, one for the training data and the other for the test data, since there were no timestamps available.

Table 1. Details of the datasets

AnyRun, HybridAnalysis			Github
Dataset type	Malicious	Benign	Benign
Training data	309	232	4901
Test data	171	92	

5.3 Environment

The experimental environment is shown in Table 2. The main libraries used to implement the experimental program are shown in Table 3.

Table 2. Experimental environment

CPU	Core i7-9700K 3.60 GHz
Memory	64 GB
OS	Windows10 Home
Programming language	Python3.7.7

Table 3. Main Python libraries used for experiments

Scikit-learn	1.0.2
Tensorflow-estimator	2.4.0
Keras	2.4.3
Keras-self-attention	0.51.0

5.4 Evaluation Metrics

The definitions of the evaluation metrics used in this study are described. Table 4 shows the relations between the predicted result and true result. True positive (TP) indicates that the system correctly evaluated a sample as malicious, and True negative (TN) indicates that the system correctly evaluated a benign sample as normal. False positive (FP) indicates that the system incorrectly evaluated a normal sample as malicious, and False negative (FN) indicates that the system incorrectly evaluated a malicious sample as normal.

Table 4. Relations between predicted and actual results

		Actual class	
		Malicious	Benign
Predicted class	Malicious	True Positive (TP)	False Positive (FP)
	Benign	False Negative (FN)	True Negative (TN)

5.5 Experiment Contents

In the experiment, we first measured the classification accuracy of the experimental model. The measurements were made by 5-fold cross validation and time series analysis. The 5-fold cross validation was performed using the entire data

set. In the time series analysis, the experimental model was trained using known samples, and the accuracy was evaluated using unknown samples.

We collected the attention weights for each word using the classification results of the time series analysis. The larger the attention weight, the stronger the influence of the corresponding word on the classification results. Therefore, we extracted the top 10 words with the highest weights from each sample and counted the number of each word.

Note that the dataset used in this study has an unbalanced sample size of benign and malicious PowerShell scripts. If such a dataset were used as it is for training the experimental model, the classification accuracy of the minority class would be reduced [6]. As a countermeasure, we applied the undersampling to our training data in the time series analysis. This is to randomly select benign PowerShell scripts, which is in the majority, in equal numbers with malicious PowerShell scripts, which is in the minority. However, the undersampling was not applied to the test data. This is because, in the actual operational environment, the probability of encountering a malicious PowerShell scripts is considerably lower than a benign one.

5.6 Result

The results of measuring the classification accuracy of the experimental model are shown in Fig. 3. The recall rate for both the 5-fold cross validation and the time series analysis exceeded 0.9, which is a high malicious PowerShell scripts detection rate. Therefore, it can be said that the experimental model has good performance as a malicious PowerShell scripts detection model. However, in the time series analysis, the F-measure decreased to about 0.7 due to an increase in the proportion of FPs.

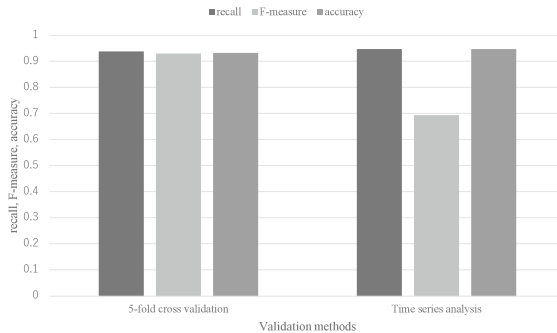


Fig. 3. Experimental model

Table 5 shows the result of the attention weights. The top 16 are excerpted for reasons of space limitation. In addition, some words have been replaced from the originals by cleartext processing. The most frequent words in the top 10 of

the samples classified as TN were commented out, followed by pipelines. Other words are considered to be commonly used for basic syntax and variable names. The most frequent words in the top 10 of the samples classified as TP were “parameter” and “position”, which are commonly used to execute functions. Also, we can see words that would be commonly used by attackers, such as “&” used for background processing and “hidden” to hide the PowerShell window. Among the words in the sample classified as FP, “dhcpcservv0scope” to obtain the scope of the DHCP server and “aesmanaged” used for encryption stand out. Many of the other words are also common to the TP results, and the reason for the misclassification can be inferred. The sample classified as FN also shows words in common with the TN results.

Table 5. Attention weight results

Rank	TN	FP	FN	TP
1	clcommentout	not	name	parameter
2	pipeline	[CLS]	add	position
3	name	static	clcommentout	function
4	[PAD]	clcommentout	clurl	@
5	server	[SEP]	dt	io
6	clurl	@	transformfinalblock	&
7	win0	dhcpcservv0scope	ciphermode	[CLS]
8	basestring	string	m	static
9	foreach	[UNK]	bytes	longstring
10	@	io	date	servicepointmanager
11	function	windows	end	[SEP]
12	password	aesmanaged	libraries	hidden
13	data	function	commit	nop
14	to	parameter	create	start
15	select	0 × 0a	cryptology	w
16	path	clurl	duck	\

6 Discussion

6.1 Important Words

The experimental results confirm certain trends for each classification result. Comparing the TN and TP columns in Table 5, very few words were common to each other, and different words were ranked by each. In addition, words in FP that were common to TP and words in FN that were common to TN occupied a large percentage of the top positions. From this, we can infer that the experimental model was misclassified by these words. Thus, the experimental model

learned from the training data that the words in the TN and TP columns of Table 5 are the features for classifying benign and malicious PowerShell scripts. Note that words such as “@” and “function” are ranked in both TN and TP. Since they indicate arrays and functions, we can infer that they are emphasized in combination with other words used at the same time in the source code.

6.2 Possibility of Evasion Attack

In this study, we were able to visualize the words that the experimental model learned from the training data and emphasized for classification. These are the important words that characterize each benign and malicious PowerShell scripts. Therefore, if a large number of words that indicate benign PowerShell script are mixed into the source code of malicious PowerShell script, we believe that there is a high possibility of misclassification. In other words, the possibility of an evasion attack would increase. However, it is important to understand such knowledge before the attacker does, in order to take countermeasures. Therefore, in addition to the results of this study, it is highly significant to conduct similar analysis on other detection models and data sets.

6.3 Research Ethics

In this study, with the help of Tajiri et al. we used the same dataset used in their study [8]. This data set is an original collection of publicly available samples. Therefore, although the number of samples is limited, the dataset is highly reproducible. In addition, the libraries and other resources used to implement the experimental model are all available free of charge, making it easy to construct the environment for this study. Therefore, we consider our study has reproducibility.

6.4 Limitations

The data set used in this study cannot be said to have a sufficient number of samples. This is because the samples were collected from sites that are freely available to everyone, with an emphasis on ensuring reproducibility. The larger the number of samples used in the experiment, the more accuracy of the analysis in this study can be expected to be improved. However, without the cooperation of security vendors and others, the number of samples is limited.

7 Conclusion

In this study, we analyzed which words in the source code characterize benign and malicious PowerShell scripts using an experimental model that combines the attention mechanism and LSTM. The experimental results visualized that samples classified as benign and samples classified as malicious were judged on the basis of different words.

Future work is to analyze using other models. There are other classification models using neural networks. It is possible that different classification models may have different distributions in the words they emphasize. By applying the same analysis to other classification models, new insights into the relationship between words and classification results may be obtained. Another future work is to evaluate evasion attacks using the results of this experiment. If a malicious PowerShell sample is mixed with words that characterize it as a benign PowerShell script, the classification model may misclassify it. We believe that it is necessary to confirm whether the classification model misclassifies or is robust by conducting experiments.

Acknowledgment. This work was supported by JSPS, Japan KAKENHI, Japan Grant Number 21K11898.

References

1. Choi, S.: Malicious powershell detection using attention against adversarial attacks. *Electronics* **9**(11) (2020). <https://doi.org/10.3390/electronics9111817>. <https://www.mdpi.com/2079-9292/9/11/1817>
2. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **18**(5–6), 602–610 (2005). <https://doi.org/10.1016/j.neunet.2005.06.042>
3. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (eds.) *Advances in Neural Information Processing Systems 21*, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 8–11 December 2008, pp. 545–552. Curran Associates, Inc. (2008). <https://proceedings.neurips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html>
4. Hendler, D., Kels, S., Rubin, A.: Detecting malicious powershell commands using deep neural networks. In: J. Kim, G. Ahn, S. Kim, Y. Kim, J. López, T. Kim (eds.) *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, AsiaCCS 2018*, Incheon, Republic of Korea, 04–08 June 2018, pp. 187–197. ACM (2018). <https://doi.org/10.1145/3196494.3196511>
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
6. Japkowicz, N.: The class imbalance problem: significance and strategies. In: *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, pp. 111–117 (2000)
7. Mimura, M., Tajiri, Y.: Static detection of malicious powershell based on word embeddings. *Internet Things* **15**, 100–404 (2021). <https://doi.org/10.1016/j.iot.2021.100404>. <https://www.sciencedirect.com/science/article/pii/S2542660521000482>
8. Tajiri, Y., Mimura, M.: Detection of malicious powershell using word-level language models. In: Aoki, K., Kanaoka, A. (eds.) *IWSEC 2020*. LNCS, vol. 12231, pp. 39–56. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58208-1_3
9. WatchGuard Technologies: Internet security report - q1 2022 (2022). <https://www.watchguard.com/wgrd-resource-center/security-report-q1-2022>. Accessed 13 July 2022



Improving Palmprint-Region Estimation for ID-Less Palmprint Recognition

Ayumi Serizawa¹, Ryosuke Okudera¹, Yumo Ouchi¹, Mizuho Yoshihira¹, Yuya Shiomi¹, Naoya Nitta², Masataka Nakahara², Akira Baba², Yutaka Miyake², Tetsushi Ohki¹, and Masakatsu Nishigaki¹(✉)

¹ Graduate School of Integrated Science and Technology, Shizuoka University, Hamamatsu, Shizuoka, Japan

`nisigaki@inf.shizuoka.ac.jp`

² KDDI Research, Inc, Fujimino, Saitama, Japan

Abstract. ID-less palmprint recognition is a biometric identification method using the pattern of the palm. It is highly available because it can obtain biometric information in a contactless manner using only a smartphone's camera in everybody's hands. It is highly convenient because it does not require an ID and uses only biometric information. Furthermore, it is highly user-friendly because the mental load on the user due to the presentation of biometric information is relatively small. On the other hand, it is an important issue for ID-less palmprint recognition to extract a palmprint region stably because the lighting environment and/or the pose of the user's hand vary, which reduces the identification accuracy. For increasing the identification accuracy in palmprint recognition, how stably the palmprint region can be extracted (referred to as the extraction stability) and the quality of the palmprint region extracted (referred to as the region quality) are important. However, the existing methods studied only extraction stability and had not been evaluated with regard to the region quality. Herein, we propose a palmprint-region estimation method using skeletal information obtained by MediaPipe that satisfies the requirements for both the extraction stability and the region quality. We investigated the palmprint region where "the differences in feature values are large between different users and small for the same user owing to differences in the pose of the hand and shooting conditions." Through our experiments, it was confirmed that the proposed method improved identification accuracy.

1 Introduction

Biometric recognition using modalities such as fingerprints, faces, and irises is now used in smartphone activations, access control, payment systems, etc. Among the biometric recognition methods using various modalities, palmprint recognition is highly available because it can obtain recognition information without contacting the user, via a camera. In addition, the mental load on the user during the presentation of biometric information is smaller than those for fingerprints and faces, making it more acceptable. Furthermore, palmprint recognition can be applied as a "show-and-go" manner that identifies a user

through only his/her palmprint. ID-less recognition, also referred as “identification”, is highly convenient because it does not require the presentation of an ID. This study focuses on ID-less palmprint recognition that satisfies the requirements of availability, acceptability, and convenience.

The procedure of enrollment and identification for ID-less palmprint recognition is as follows. In the enrollment phase, the user’s palm is captured using a camera. The palmprint region, a.k.a. Region of Interest (referred to as ROI), is extracted from the palm image and enrolled as a ROI image (template image). In the identification phase, the user’s palm is captured and the palmprint region is extracted to obtain a ROI image (query image). The query image is compared with the template images of all enrollees. The user is identified when a template image is found that is sufficiently similar to the query image.

Two major issues remain in the palmprint recognition method. The first is the long identification time for finding the legitimate user who presented the query image among all the enrollees. As the number of enrollees increases, the amount of time needed to find a template image that matches the query image presented increases. The second is the low identification accuracy for palm images obtained in various environments and conditions. If the lighting environment or the pose of the hand differs each time an image is obtained, it is difficult to extract the palmprint region stably, which reduces the identification accuracy.

Google Inc. Released MediaPipe—a framework using machine learning with multimedia data [4]. A set of artificial-intelligence models is included in MediaPipe, which can detect faces and estimate the skeletal structures of bodies and palms in images and videos. Among them is MediaPipe (Hands). By using this model, it is possible to estimate the palm skeleton in palm images and obtain 21 coordinates (skeletal information), as shown in Fig. 1. In palmprint recognition, a palm image is captured to obtain a ROI image. Skeletal information can be extracted from this palm image using MediaPipe (Hands). This is why palmprint recognition and MediaPipe are highly effective when combined. We aim to tackle the above-mentioned issues of the palmprint recognition by using skeletal information.

In this study, we focused our objective on to improve the second issue, i.e., increase the accuracy of palmprint-region estimation. To achieve stable palmprint recognition, preprocessing is essential for correctly estimating the palmprint region from the palm image. For palmprint-region estimation, both the extraction stability, i.e., how stably the palmprint region can be extracted from the palm image, and the region quality, i.e., how appropriately the palmprint region can be chosen, are important.

Several methods have been proposed to improve the extraction stability. For example, a method for estimating the palmprint region by detecting the lines between fingers was proposed by Agematsu et al. [1]. However, depending on the condition of the palm image, there are cases where the interdigital lines cannot be detected. In such cases, the palmprint region cannot be estimated either. Then, an alternative method for estimating the palmprint region using skeletal information of the palm was proposed by Nitta et al. [2]. This method significantly improves the extraction stability. However, the region quality is not addressed sufficiently. In our study, we attempted to improve the region quality of Nitta et al.’s palmprint-region estimation method (hereinafter referred to as the

existing method). Specifically, we investigated regions in palm images where differences between individuals are large and variations within individuals are small.

We developed three palmprint-region estimation methods and compared them with the existing method. Palm images from 523 people (10 images for each person) were used to evaluate the proposed methods. The results confirmed that the proposed methods achieved higher identification accuracy in ID-less palmprint recognition than the existing method.

2 Related Work

2.1 Palmprint-Region Estimation Using Hand Shape

Oldal et al. proposed a method for estimating the palmprint region according to the hand shape [3]. First, the palm image is preprocessed in the following order: grayscaling, segmentation, denoising, and edge detection. Fingertips are detected based on the obtained edges. The finger valleys (bases of the fingers) are detected using the detected fingertips, and the palmprint region is estimated according to the points of the finger valley. The method assumes that a palm image with a dark background and bright hand is obtained. An image with a dark background and bright hand has a bimodal distribution; thus, the background can be separated from the hand by setting an appropriate threshold in the segmentation process. However, in actual scenes of smartphone use, various backgrounds may appear in the image. When the palm image does not have a bimodal distribution, it is difficult to estimate the palmprint region.

2.2 Palmprint-Region Estimation Using Interdigital Spaces

Agematsu et al. proposed a method for palmprint-region estimation based on interdigital spaces [1]. The method detects the index, middle, ring, and little fingers in a palm image and estimates the palmprint region using the line segments between the fingers. Specifically, three line-segments between the index, middle, ring, and little fingers are detected via line detection processing of a palm image with the fingers closed. Next, the method estimates the finger valleys according to the three detected line segments. Finally, the method estimates the palmprint region using the estimated finger valley points (Fig. 1). However, in actual scenes of smartphone use, users may not have their fingers closed. If the fingers are open, the line segments between the fingers cannot be extracted correctly, and the palmprint-region estimation fails.

2.3 Palmprint-Region Estimation Using Skeletal Information

Google Inc. Released MediaPipe—a framework using machine learning with multimedia data [4]. They also released a set of artificial-intelligence models using MediaPipe. Among them is MediaPipe (Hands), which is a model for hand tracking and can obtain 21 points of skeletal information from a video or image of a hand (Fig. 2 (left)).

By using skeletal information obtained by MediaPipe (Hands), Nitta et al. proposed two methods for estimating the palmprint region from a palm image (referred to as existing methods 1 and 2) [2]. Existing method 1 uses two points of the skeletal information:

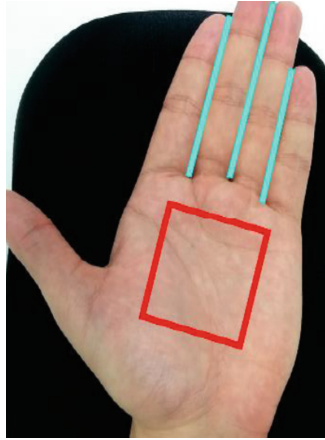


Fig. 1. Example of palmprint-region estimation for method [1]

the base of the index finger (point I) and the base of the little finger (point L). The middle point of the line segment IL is A, and a couple of points I' and L' are defined on the line segment IL such that $I'A + AL' = 0.9IL$. A square region on the palm with the line segment I'L' as one side (white square in Fig. 2 (middle)) is extracted. The square region is resized to obtain a 160×160 [px] square ROI image (Fig. 2 (upper right)). Existing method 2 uses four points of the skeletal information: the base of the index finger (point I), the base of the little finger (point L), the base of the thumb (point T), and the wrist (point W). A point T' is defined on the line segment IT, such that $IT' = 0.9IT$. Similarly, a point W' is defined on the line segment LW' = $0.9LW$. Then, the square ILW'T' (green square in Fig. 2 (middle)) is perspective-transformed to obtain a 160×160 [px] square ROI image (Fig. 2 (lower right)).

These methods are based on MediaPipe, which can stably estimate skeletal information. Thus, they have good extraction stability. However, the region quality was not considered in the study of Nitta et al.

3 Proposed Methods

3.1 Improvement of Region Quality

For increasing the accuracy of palmprint recognition, both the extraction stability and the region quality are important. For the extraction stability, improvements have been made by researchers, as described in Sect. 2. In contrast, to our knowledge, the region quality has not been sufficiently considered so far. In this study, we attempted to improve the quality of the palmprint region for existing method 1 [2] described in Subsect. 2.3. Specifically, we investigated regions in palm images where differences between individuals are large and variations within individuals are small. When we observe our palms, we see that the variations in the palmprint caused by finger movements are more significant closer to the bases of the fingers. In the region near the base of the thumb, the skin of the palm expands with finger movements, causing significant palmprint variations.

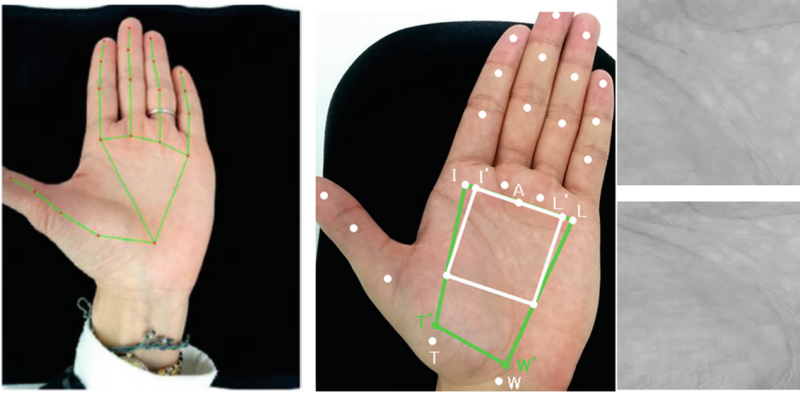


Fig. 2. Example of palmprint-region estimation for two methods [2] (left: skeletal information; middle: palmprint-region estimation; upper right: ROI image (white); lower right: ROI image (green))

In short, the base of the fingers (particularly the thumb) is a highly variable region in the palm of the hand. Therefore, from the aspect of variations within individuals, the regions around the fingers should not be included in the palmprint region used for palmprint recognition. In this paper, we propose three palmprint-region estimation methods based on existing method 1. The first method (proposed method 1) aims to improve the region quality by avoiding the base of the thumb; the second method (proposed method 2) is similar but avoids the base of all five fingers; the third method (proposed method 3) is a simpler version of the proposed method 1. Figure 3 shows examples of palmprint-region estimation for these methods, and Fig. 4 shows examples of palmprint regions estimated via these methods.

3.2 Proposed Method 1

The first method aims to avoid the region where existing method 1 has the biggest problem, i.e., the region near the base of the thumb. The proposed method 1 uses the points I' and L' defined in existing method 1. A square region on the palm with line segment $I'L'$ as one side is extracted. The vertices of this square (excluding points I' and L') are called points B and C. A point I'' is defined on the line segment $I'L'$, such that $I''L' = 0.8I'L'$. A point C' is defined on the line segment $L'C$, such that $L'C' = 0.8L'C$. Then, a square region with the line segments $I''L'$ and $C'L'$ as two sides is extracted (blue region in Fig. 3). This square region is resized to obtain a 160×160 [px] ROI image (Fig. 5(b)).

3.3 Proposed Method 2

The second method aims to avoid the regions near the base of the index, middle, ring, and little fingers in addition to the region near the base of the thumb. The proposed method 3 uses the points I' , L' , B, and C, too. A point B' is defined on the line segment

BC, such that $B'C = 0.8BC$. A point L'' is defined on the line segment $L'C$, such that $L''C = 0.8L'C$. Then, a square region with the line segments $L''C$ and $B'C$ as two sides is extracted (red region in Fig. 3). This square region is resized to obtain a 160×160 [px] ROI image (Fig. 5(c)).

3.4 Proposed Method 3

The third method aims to propose a simpler method for a better region estimation. The proposed method 3 uses two points of the skeletal information: the base of the middle finger (point M) and the base of the little finger (point L). A square region on the palm side with the line segment ML as one side is extracted (yellow region in Fig. 3). This square region is resized to obtain a 160×160 [px] ROI image (Fig. 5(d)).

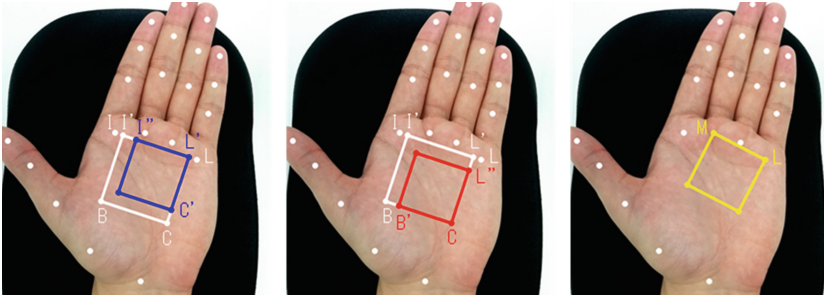


Fig. 3. Proposed methods for palmprint-region estimation (left: proposed method 1; middle: proposed method 2; right: proposed method 3)

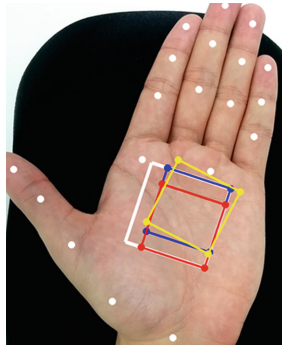


Fig. 4. Palmprint regions estimated via each method (white region: existing method 1; blue region: proposed method 1; red region: proposed method 2; yellow region: proposed method 3). (Color figure online)

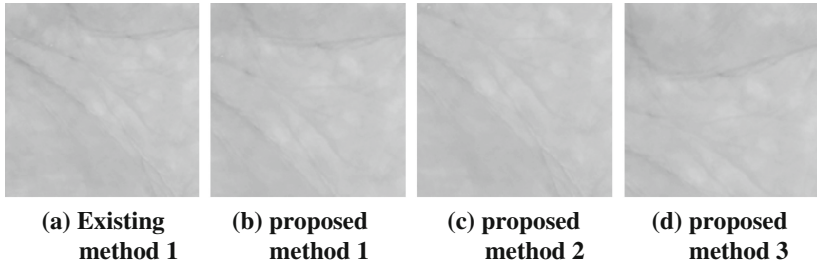


Fig. 5. ROI images for each method

4 Evaluation

4.1 Evaluation Dataset

10 palm images were obtained from each of 523 people and used to evaluate each method. In this experiment, the left palms of all the users were captured in the same environment. An example of palm image in our experiment is shown in Fig. 6.



Fig. 6. Example of palm image

4.2 Evaluation Method

A palm image for enrollment and a palm image for identification were selected at random from the 10 palm images for each user. Existing method 1 and proposed methods 1–3 were applied to the 523 palm images for enrollment and 523 palm images for identification to obtain ROI images for enrollment (template images) and ROI images for identification (query images). The concrete procedures of the process are as follows. First, each palm image was fed to MediaPipe (Hands) to extract skeletal information. Then, the palmprint region was estimated by applying existing method 1, proposed method 1, 2, and 3, respectively. Finally, each estimated region was resized to obtain a 160×160 [px] ROI image.

To calculate the false rejection rate (FRR), the verification process described in Sect. 4.3 was applied to all the combinations of template images and query images for the same user (523 combinations). Additionally, to calculate the false acceptance rate

(FAR), this verification process was also applied to all the combinations of template images and query images for different users (523×522 combinations). Then, the equal error rate (EER) was calculated by determining the matching threshold where the FRR was equal to the FAR and was used as a measure of the identification accuracy. The above procedure was repeated ten times, and the average EERs for each method were compared.

4.3 Verification Process

We applied the verification process proposed in [5], which has been used in previous studies [1][1].

1. A template image (160×160 [px]) and a query image (160×160 [px]) were inputted to be compared. These images were referred as a high-resolution (HR) template and a HR query, respectively.
2. The HR template was averaged every 2×2 [px] to generate 80×80 [px] medium-resolution (MR) template. The MR template was further averaged every 2×2 [px] to generate 40×40 [px] low-resolution (LR) template.
3. The HR query was averaged every 2×2 [px] to generate 80×80 [px] medium-resolution (MR) query. The MR query was further averaged every 2×2 [px] to generate 40×40 [px] low-resolution (LR) query.
4. Band-limited phase-only correlation (BLPOC) with a window size of 32 was applied to the LR template and the LR query to calculate the amount of displacement between the template and query.
5. The MR template and the MR query were divided into 16 blocks of 20×20 [px]. In this process, the query was aligned by the amount of displacement calculated in step 4.
6. BLPOC was applied to the first blocks of the MR images generated in step 5 to calculate the amount of displacement between them. This process was repeated for all the blocks to calculate the amount of displacement between the MR template and the MR query for each block.
7. The HR template and the HR query were divided into 16 blocks of 40×40 [px]. In this process, the query image was aligned for each block by the amount of displacement calculated in step 6.
8. BLPOC was applied to the first block of the images generated in step 7 to calculate the amount of displacement between them. Then, a calculation was performed to determine which point in the first block of the HR query corresponds to the center of the first block of the HR template. The center point and the corresponding point were recorded as a corresponding point pair only when the matching score obtained using BLPOC exceeds 0.3. This process was repeated for all the blocks to record the corresponding point pairs for each block.
9. Affine transformation was applied to the HR query using all the corresponding point pairs recorded in step 8.
10. The central region (96×96 [px]) of the HR template was extracted and divided into nine blocks of 32×32 [px].

11. The central region (96×96 [px]) of the HR query obtained in step 9 was extracted and divided into nine blocks of 32×32 [px].
12. BLPOC was applied to the first block of the HR template generated in step 10 and the first block of the HR query generated in step 11 to calculate the amount of displacement between them. Then, a calculation was performed to determine which point of the first block of the HR query corresponds to the center of the first block of the HR template. This process was repeated for all the blocks to record the corresponding point pairs of each block.
13. Considering the amount of displacement for each block calculated in step 12, the central region (96×96 [px]) was extracted from the HR query obtained in step 8 and divided into nine blocks of 32×32 [px].
14. BLPOC was applied to the first block of the HR template generated in step 10 and the first block of the HR query generated in step 13 to obtain the BLPOC image. The pixel value of the coordinates (i, j) in the BLPOC image shows the matching score for the first block of the HR template and the HR query, where the HR query is shifted by i pixel in the x direction and j pixel in the y direction. These processes were repeated for all the blocks to obtain a total of nine BLPOC images (32×32 [px]).
15. An averaged image of the nine BLPOC images was generated. The maximum pixel value in the averaged image was output as the matching score for the HR template and the HR query input in step 1.

4.4 Results

The FAR and FRR for each method are shown in Fig. 7. The EER for each method is presented in Table 1. As mentioned in Sect. 4.3, the verification procedure described in Sect. 4.3 was repeated ten times. Figure 7 shows graphs of the first FAR and FRR. Table 1 presents the average and standard deviation of all ten EERs.

4.5 Discussions

As indicated by Table 1, the EER of existing method 1 was greater than those of proposed methods 1 and 3. We considered that regions included only by existing method 1 but not by proposed methods 1 and 3 decreased the identification accuracy. We investigated the ROI images for which existing method 1 failed to identify the user. We found that the matching scores for the ROI images showing a wrinkle at the base of the thumb (red region in Fig. 8) were relatively low. This region was included only by existing method 1 and tended to be easily variable for multiple images from the same user. Hence, from the perspective of the region quality, we can say that this region is not suitable for palmprint recognition.

Table 1 shows that the EER of proposed method 2 was greater than those of proposed methods 1 and 3. We considered that regions excluded by proposed method 2 and included by proposed methods 1 and 3 increased the identification accuracy. The region excluded by proposed method 2 corresponded to region (1) in Fig. 9. In this region, there were intelligence line and emotion line of palmistry. It is known that the lines of palmistry differ among different users. Therefore, this region tended to have larger

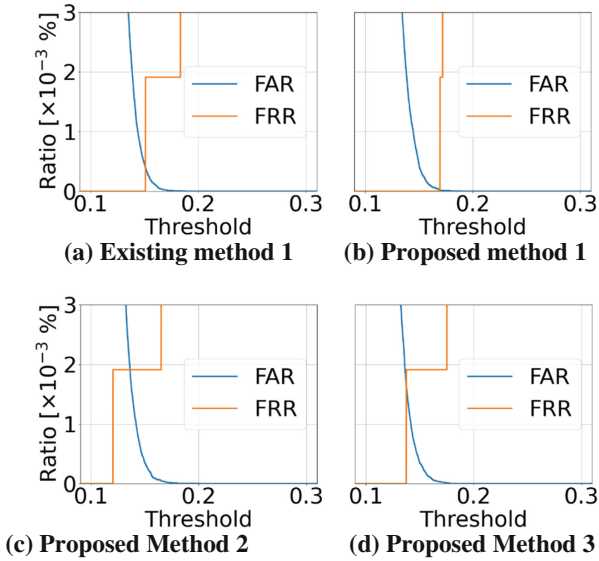


Fig. 7. FAR and FRR for each method

Table 1. EER for each method

Method	Average EER [%]	Standard deviation
Existing method 1	1.52×10^{-3}	1.24×10^{-3}
Proposed method 1	3.85×10^{-4}	7.64×10^{-4}
Proposed method 2	1.68×10^{-3}	4.85×10^{-4}
Proposed method 3	5.42×10^{-4}	8.31×10^{-4}

differences for each user. Hence, from the perspective of the region quality, we can say that this region is suitable for palmprint recognition.

Furthermore, Table 1 shows that the EER of proposed method 1 was lower than that of proposed method 3. We considered that the regions excluded by proposed method 1 but included by proposed method 3 increased the identification accuracy. The corresponding regions are regions (2) and (3) in Fig. 9. Region (2) is near the base of the middle, ring, and little fingers, and in this region, the variations of the palmprints in response to the finger pose were large. Region (3) is near the base of the thumb, and in this region, the palmprints varied depending on the pose of the base of the middle and little fingers. Hence, from the perspective of the region quality, we can say that these regions are not suitable for palmprint recognition.

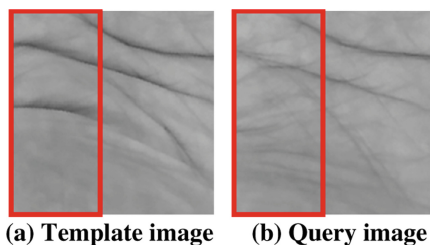


Fig. 8. Example pair of ROI images with low matching score

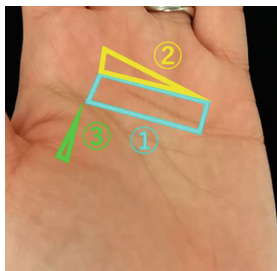


Fig. 9. Unique regions included only in proposed methods 1 and 3

5 Conclusion

We increased the identification accuracy for palmprint recognition by improving the region quality in palmprint-region estimations. Specifically, the regions of the palm with the following two characteristics were considered.

- The differences among users are large.
- Within the same user, the variations due to differences in the finger pose and the environment for capturing the images are small.

Palmprint variations in response to finger movements are relatively large near the base of the fingers (particularly the thumb). Accordingly, we proposed three methods to improve the palmprint-region estimation used in existing method 1 [2]. The accuracies of these methods were evaluated and compared, and our methods achieved higher identification accuracies than existing method 1 for palmprint recognition. The following problems were found by reviewing the ROI images.

- The wrinkles at the base of the thumb are affected by variations in the thumb pose. If a region includes near the base of fingers (especially the thumb), the region has variation within the same user. This reduces the identification accuracy.
- If a region includes palmistry lines in the ROI image, the region has differences among users. This increases the identification accuracy.

In addition to the palmprint region, the identification time is a problem in palmprint recognition. In the future, we will investigate methods to further improve palmprint recognition from various viewpoints using skeletal information.

References

1. Agematsu, H., Kanda, R., Matsui, T., Miyake, Y., Ito, K., Aoki, T.: Palmprint recognition using interdigital spaces line. In: Computer Security Symposium (CSS 2016), vol. 2, pp. 185–191 (2016). (in Japanese)
2. Nitta, S., Nakahara, M., Baba, A., Miyake, Y.: A method for estimating palmprint regions using skeletal information in palmprint authentication. In: Symposium on Cryptography and Information Security (SCIS 2021), vol. 3F3–2, pp. 1–6 (2021). (in Japanese)
3. Oldal, L.G., Kovács, A.: Hand geometry and palmprint-based authentication using image processing. In: 2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY). IEEE, (2020)
4. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: MediaPipe hands: on-device real-time hand tracking. arXiv preprint [arXiv:2006.10214](https://arxiv.org/abs/2006.10214) (2020)
5. Aoyama, S., Ito, K., Aoki, T., Ota, H.: A contactless palmprint authentication algorithm for mobile phones. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **J96-A**, 250–263 (2013). (in Japanese)



Real Vehicle-Based Attack Dataset for Security Threat Analysis in a Vehicle

Yeji Koh, Yoonji Kim, Munkhdelgerekh Batzorig, and Kangbin Yim^(✉)

Department of Information Security, Soonchunhyang University, Asan, South Korea
{julysnowflake,rladbsw17,munkhdelgerekh,yim}@sch.ac.kr

Abstract. Modern Connected and Autonomous Vehicles (CAVs) are equipped with an increasing large number of Electronic Control Units (ECUs). These ECUs transmit Control Area Network (CAN) protocol in data exchange via an in-vehicle network. Since CAN message uses broadcast transmission, it becomes hinder to detect the intrusion. Thus, building an intrusion detection system (IDS) and intrusion prevention system (IPS) to IVN is essential, and most effective method for those system is effective usage of deep learning (DL) or machine learning (ML); Nevertheless, building an IDS or IPS based on DL or ML is required huge amount of data of normal state and during the attack. Consequently, there is a high demand for an accurate normal and attack dataset for IVN protection. Therefore, we propose accurate environment configuration for attack data collection from the vehicle.

1 Introduction

As the development of the modern automobile industry, the number of electronic control units (ECUs) that maintains the vehicle has increased rapidly, increasing the complexity of communication interfaces. Controller Area Network (CAN) protocol is more efficient than Conventional Universal Asynchronous Receiver/Transmitter (UART) methods. Thus, CAN protocol is the most frequently used protocol in IVN for transmitting and receiving data from ECUs. Since CAN protocol uses broadcast transmission as to communicate with other nodes, it is hinder to apply an authentication mechanism to IVN. Thus, if an attacker unauthorized access through the CAN bus and manipulates data, it is impossible to distinguish between the error message and the actual message, which can lead a great risk [1]. The possible attack models through the CAN bus are same as follows:

Injection Attack. Through OBDII and ECU systems, it enters the network inside the vehicle and attacks by injecting attack data. The CAN network cannot identify whether the received frame is legitimate because it does not have an authentication mechanism.

DoS (Denial of Service) Attack. Since the CAN protocol uses broadcast transmission, there is no source or destination address, and all messages are received synchronously to one ECU. To avoid this collision, CAN protocol uses ID field to calculate priority of order of the messages. If message has low ID value, it represents a higher priority, if there

is a frame with the highest priority on the CAN bus, other nodes cannot send message frames to the IVN until high priority message is served. Thus, the attacker always sets the highest priority frame (0x00) to be transmitted to the in-vehicle network to attack in a form in which other nodes are placed in the standby state [2].

Fuzzing Attack. Fuzzing attacks are similar to indiscriminate attacks in the form of randomly injecting ID values and data into CAN networks. Attack data can be easily injected using the characteristics of the CAN network, where authentication is not performed on the transmission and reception nodes, which can cause an unexpected situation while driving the vehicles [3].

Replay Attack. The attack is executed when the vehicle injects a message that causes a specific function or behavior different from the current state into the message cycle of the existing state [4].

Spoofing Attack. A spoofing attack is a hacking attack that an adversary approaches on a network pretending to be an authorized address of the system to gain unauthorized access [5].

The ECU transmits data by sending and receiving CAN messages through the CAN network. If adversary such as Dos, Bus-off, and Fuzzing are activated due to the vulnerability of the CAN, data on the CAN cannot be read or written, resulting in serious consequences that damage to the ECU or even threaten the safety of the driver. In this paper, using the vulnerability of CAN, we select a total of three attack models for real vehicles: DOS, Fuzzing, and Replay to show the effect on the vehicle injected with the attack message, and collect attack datasets generated through the attack. The collected attack datasets can be used to develop Intrusion Detection System (IDS) for vehicles and machine learning for vehicles to prevent security threats in preparation for future attacks on actual vehicles. The paper is organized as follows. Section 2 explains similar research works about possible attacks on IVN. Section 3 explains methods and configuration of environment for CAN attack datasets collection, and tools for inject data into vehicle internal networks. Section 4 provides detailed reactions of the vehicle when data is injected. Finally, the conclusions and future works are drawn in Sect. 5.

2 Related Work

In the vehicle, various ECUs are integrated into the CAN bus system to exchange CAN messages. CAN does not apply security for various attacks, so several vulnerabilities are found, and Tianxiang Huang et al. analyze attack models for CAN buses and design and develop ATG (Attack Traffic Generation Tool) to explain key features. ATG can define the attack types, content, set timer-based execution and build an effective function GUI using wxPython library. Types of attacks include Dos, Fuzzing, Spoofing, etc. DoS attack send messages every 0.8 ms from 2 s to 4 s, and a series of random messages from 8 s to 40 s for fuzzing attacks. ATG that the proposed method can share and reuse attack scenarios for CAN bus security testing and supports CAN database conversion capabilities.

In addition, it is possible to migrate to CAN networks as well as other network communication protocols such as CAN-FD, Flex-Ray, and Ethernet. The authors construct the attack via connection with On-Board Diagnostics (OBD) using real cars and inject CAN messages directly to show the speedometer decreasing to zero. We confirm that the recently released OBD ports of vehicles have a limited type of CAN network used for analysis and attack, and in this paper, we test and prove whether attack injection for various CANs is possible using the EDA (CAN gateway ECU Direct Approach) method [6].

Kazuki Iehira et al. propose a method of transitioning to the bus-off state by intentionally setting the message transmitted by the ECU to an error based on a spoofing attack method in which the receiving and transmitting ECU does not detect anomalies in the applied ECU using a bus-off attack. The CAN message is a broadcasting method, and since there is no transmission address and only the CAN ID exists as the destination address, spoofing attacks are easily possible. Three methods are presented, and the results are shown: bus-off attack using bit errors, bus-off attack using stub errors, and bus-off attack using one frame. This paper consisted of attack hardware and target ECU for CAN bus experiments and used Field-Programmable Gate Array (FPGA) to inject spoofing messages. The spoofing target was the engine speedometer of a stationary vehicle, and in the case of simple spoofing, the speedometer's pointer fluctuates, but the proposed method of spoofing using bus-off attacks in this paper did not detect errors and no anomalies in the vehicle used in the experiment. They validate the attack in a simulated environment consisting of attack hardware and ECU and evaluate the impact of spoofing attacks effectively implemented on real cars [7].

Cyber-attacks on vehicles are likely to cause casualties. Shahida Malik and Weiqing Sun conduct an analysis of cyberattacks on connected and autonomous vehicles and simulate cyberattacks based on the analysis to show the actual damage impact. They used CARLA 0.9.6 for simulation and controlled the environment differently depending on the scenario. They simulated a total of 10 significant and high-priority attack scenarios and recorded important functions, including collision, position at intersections, speed, gear, position, orientation, and acceleration. They apply scenario that wireless update attacks, cyber terror attacks, mobile app attacks, OBD dongle attacks, etc., it shows results such as stop the car engine or potentially stop the car moving through Bluetooth and disable emergency support services. We confirm that these attacks are fully simulated, and we evaluate the quality of the attack dataset by injecting and testing attack messages in direct-driving vehicles, verifying their impact on the vehicle, and collecting attack datasets [8].

3 Method of Generate Attack CAN Dataset Based on Actual Vehicle

3.1 Method for Collection CAN Dataset in Vehicle

In this paper, we used to EDA (CAN gateway ECU Direct Approach) method to collect CAN dataset in vehicle [9]. EDA is method to access the CAN gateway ECU through ECU's tapping line and collect actual vehicle internal network data by using CAN dataset collection tools (Fig. 1).

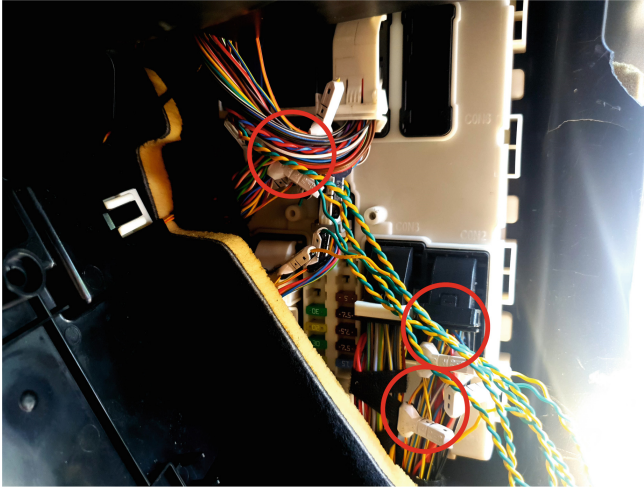


Fig. 1. Shows the actual vehicle tapping into the CAN gateway module using the EDA method. The CAN network connection is for K-CAN, K-CAN2 and PT-CAN, and each CAN is responsible for body control, multimedia function control, etc.

Mostly Much internal network research of vehicle uses OBDII port to access the network. However, because there are located just the diagnostic CAN data in OBDII port, it hard to check entire CAN bus of the vehicle. So, limits of check and research to diverse vehicle functions always exist.

On the other hand, the EDA method can collect more diverse function data of vehicle than collection data using OBDII port. By using this method, we could collect more accurately performed of detailed analysis and attack to target CAN data for a specific function. Also, it is possible to collect more accurate and larger amount of data than the data of OBDII port because EDA method can access to entire CAN bus network. In terms of vehicle internal network research, this method makes the result more effective and accurate than previous research.

3.2 Vehicle Internal CAN Network Attack Tool

In order to analyze the threat of the vehicle's internal network, high-quality attack data is required for researchers. The high-quality attack data means data that is difficult to detect because the attack data is similar to the actual driving data. If the attack data is similar to the actual driving data, it is difficult to find anomalies in contrast to the actual data, making it difficult to detect them with Intrusion Detection System (IDS).

To improve the existing IDS of the vehicle, we developed more sophisticated internal network attack tools to generate effective attack data. In this paper, we implement attack tools against Fuzzing, DoS, and Replay attacks. Figure 2 below is a schematic diagram of the attack process by injecting an attack data packet into an actual vehicle.

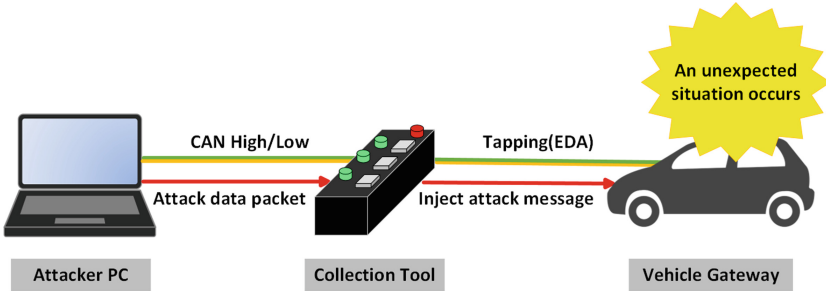


Fig. 2. Actual vehicle-based attack dataset injection process

In this paper, we developed an internal CAN network attack tool based on the characteristics of each attack. The purging attack tool is set to inject arbitrary data periodically, allowing attackers to specify the range of arbitrary data values, the range of IDs, the number of injections, and the length of arbitrary data. If the attacker runs the purging tool, a random dataset is injected and malfunctioning.

When the DoS attack tool is executed, a large number of high-priority IDs are injected within a short time. As a result, the operation of the data packet having a low priority is blocked.

3.3 Collection of Attack CAN Data Packet Based on Actual Vehicle

Using our own developed attack tools, we collected attack CAN data packets based on real vehicles. In order to collect the IVN attack data set of the real vehicle, line tapping was performed on the CAN gateway module using the EDA method. When collection begins, the vehicle's CAN data packet is collected through the tapping line of the CAN gateway module. In addition, the attack data packet is injected into the CAN network of the real vehicle using the tapping line.

The CAN network attack based on the actual vehicle is performed during driving. In the case of a purging attack, basic functions and eventuality functions are activated through random value data packet injection. Dos attacks inject attack data packets to activate certain eventuality features.

In this study, attacks were carried out during driving on CAN networks mainly related to vehicle control and visual parts. In preparation for unexpected situations, attacks on body-related functions took place only at rest. If an attacker attacks a large amount of data while collecting driving data packets, a collision between the collection data and the attack data may occur. In this case, an appropriate attack packet may not be generated. Therefore, to avoid this problem, we made a scenario in which attacks are performed at regular intervals for a short time.

The attack scenario is as follows (Table 1):

Table 1. Actual vehicle-based CAN internal network attack scenario

Attack	Scenario
Fuzzing	- Inject 100 data every 2 min for 3 s - Inject 500 data every 2 min for 3 s - Inject 100 data every 2 min for 5 s - Inject 500 data every 2 min for 5 s
DoS	- Inject 5,000 data every 2 min for 2–3 s - Inject 10,000 data every 2 min for 2–3 s
Replay	- Replay pre-collected injection data for 5–6 s

All attack scenarios were carried out only in certain sections where the risk of accident rate was low. To prevent collision between each data packets, we use to scenarios in which 100 to 500 data are injected for 3 to 5 s at 2-min intervals in Fuzzing attack. In case of DoS attack, 5,000 to 10,000 data inject for 2 to 3 s at 2-min intervals. And Replay attack inject pre-collected data for 5 to 6 s at 5-min intervals. By injecting CAN attack data packets based on actual vehicle, the vehicle’s functions such as emergency light and warning light on the instrument panel were activated. And function that could be contacted by the manufacturer in an emergency situation was successfully attacked (Fig. 3).



Fig. 3. Perform RDC initialization via fuzzing attack on real vehicle

Collecting CAN network attack data from real-world vehicles allows us to identify potential attack risks in real-world situations rather than data collected from virtual environments or stationary vehicles. You can also attack features that cannot be performed on a simple test bed. It is possible to analyze data related to more functions by analyzing data collected during an actual vehicle attack. And it can have an effective impact on security threat detection and analysis studies on vehicle interior networks.

4 Evaluation

In this section, we describe the reactions of vehicle functions when we inject the attack datasets into the vehicle. We were able to develop our own program for each attack using P-Code library that provides from peak-system [10].

4.1 Result of DoS Attack

The nature of DoS attack is slow down or shut down a functions or network, making it inaccessible to its intended node. We accomplished this attack by flooding IVN with high priority ID message which is 0x00 ID. As result, we were able to check functions like frequency of signal light, dashboard information etc., are slowing down, and could check time offset of messages that collected from the vehicle.

```

from PCANBasic import *
import time
def DoS_Attack():
    time_offset = 0.03
    DoS_attack_data = (0x00, 0x04, 0x81, 0x00, 0xC0, 0x02,
0x20, 0x4D)
    DoS_attack = TPCANMsg()
    DoS_attack.ID = int(0x000)
    DoS_attack.LEN = 8
    DoS_attack.MSGTYPE = PCAN_MESSAGE_STANDARD
    DoS_attack.DATA = DoS_attack_data
    for i in range(0, 10000):
        CAN.Write(CAN_BUS, DoS_attack)
        time.sleep(time_offset)
CAN = PCANBasic()
CAN_BUS = PCAN_USBBUS2
CAN.Initialize(CAN_BUS, PCAN_BAUD_500K, 2047, 0, 0)

while True:
    DoS_Attack()

```

4.2 Result of Fuzzing Attack

The main purpose of fuzzing attack is to exploit the possible vulnerabilities or function messages. We initiated fuzzing attack by injecting messages with random ID and Data. We have developed program for this attack. With this attack, we were able to inject data as scenario follows, and same time we could save injected data into another text file for data labeling. As mentioned in Sect. 3, we collected attack dataset while driving in highway, and at the same time, we had injected data into the vehicle, because of this, we carried out the risk that threatening to driver's life, because of the risk, we have only injected data to K-CAN which mainly assigned for controllers of the vehicle's dashboard. After the fuzzing attack, we found following function's data, and store in excel file (Table 2).

Table 2. Vehicle function findings after fuzzing attack

Function	Description
Speed in dashboard	Change the speed measurement in dashboard
RPM	Change RPM measure in dashboard
Gear (PARK)	Change the gear to parking mode
Gear (Drive)	Change the gear to drive mode
Signal Light (Left)	Change the status of turn left signal light to on
Signal Light (Right)	Change the status of turn right signal light to on
Emergency Light	Change the status of emergency light to on
Open door	Open all doors
Front ultra-sonic wave	Give a signal of detected to radar sensor. (Give Max, Medium and Low)
Back ultra-sonic wave	Give a signal of detected to radar sensor. (Give Max, Medium and Low)

4.3 Result of Replay Attack

Main intention of replay attack is fraudulently delays or resends it to misdirect the receiver into doing what the adversary wants. Thus, adversary needs information about the system, to initiate a replay attack. After injecting fuzzing attack successful, we have obtained some message formats that assigned for certain functions. Based on this information, we could initiate replay attack by injecting those messages, and control vehicle's some functions. For replay attack on the vehicle, we have developed attack tool with user interface using P-CAN library of python. Replay attack program has main two roles. The first one is to inject the data from list that gained functional messages after fuzzing attack. The other one is to save injected time of messages for later data labelling (Fig. 4).

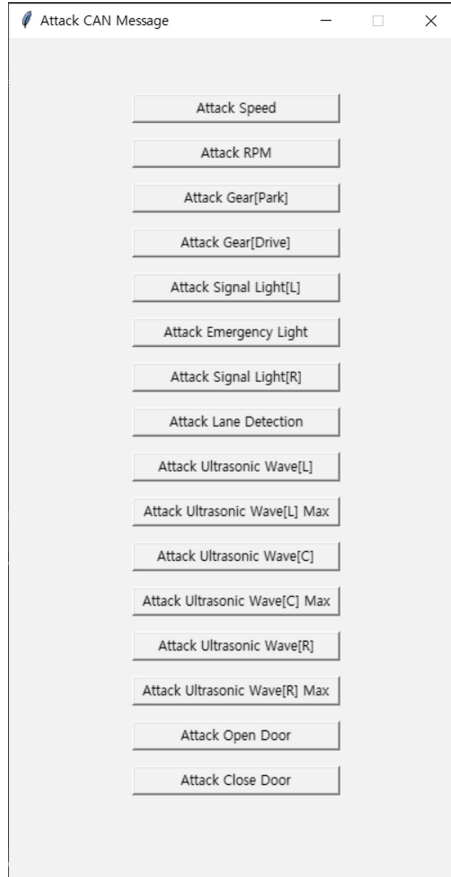


Fig. 4. Display of the tool for replay attack to IVN

5 Conclusion and Future Work

In modern automobiles, the number of ECUs mounted inside the vehicle is increasing exponentially for the safety and convenience of drivers. However, the need for vehicle security has also increased as vehicle attack methods with serious consequences that threaten the driver's life continue to expand. Therefore, an accurate analysis of the type of attack that occurred is required, and it is important to recognize that there is a high possibility of an attack on the vehicle, and to anticipate and prepare for an attack on the vehicle.

In this paper, DoS, Fuzzing, and Replay attacks were performed on actual vehicles driving with 7 scenarios. As a result of the experiment, the vehicle window was opened, the emergency SOS button was activated, and navigation reboot etc., this resulted in an unexpected situation in which the driver could be embarrassed while driving. In addition, we collected CAN data from the ECU by selecting the existing EDA method and made attack tools using the characteristics of each attack. In real-world vehicles,

attack messages were injected into the CAN bus to generate high-quality attack datasets that can later be used as materials for vehicle security. In the future, we will apply our attack dataset into IDS or IPS system, and compare performance with other public attack dataset.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A2001810) and This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01343, Regional strategic industry convergence security core talent training business).

References

1. Hong, S.: Research on countermeasures of controller area network vulnerability. *J. Converg. Inf. Technol.* **8**(5), 115–120 (2018)
2. Biron, Z.A., Dey, S., Pisu, P.: Real-time detection and estimation of denial of service attack in connected vehicle systems. *IEEE Trans. Intell. Transp. Syst.* **19**(12), 3893–3902 (2018). <https://doi.org/10.1109/TITS.2018.2791484>
3. Mukherjee, S., Shirazi, H., Ray, I., Daily, J., Gamble, R.: Practical DoS attacks on embedded networks in commercial vehicles. In: Ray, I., Gaur, M.S., Conti, M., Sanghi, D., Kamakoti, V. (eds.) *ICISS 2016. LNCS*, vol. 10063, pp. 23–42. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49806-5_2
4. Chandrasekaran, S., Ramachandran, K.I., Adarsh, S., Puranik, A.K.: Avoidance of replay attack in CAN protocol using authenticated encryption. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6 (2020). <https://doi.org/10.1109/ICCCNT49239.2020.9225529>
5. Yang, Y., Duan, Z., Tehranipoor, M.: Identify a spoofing attack on an in-vehicle CAN bus based on the deep features of an ECU fingerprint signal. *Smart Cities* **3**, 17–30 (2020). <https://doi.org/10.3390/smartcities3010002>
6. Huang, T., Zhou, J., Bytes, A.: ATG: an attack traffic generation tool for security testing of in-vehicle CAN bus. In: *Proceedings of the 13th International Conference on Availability, Reliability and Security* (2018)
7. Iehira, K., Inoue, H., Ishida, K.: Spoofing attack using bus-off attacks against a specific ECU of the CAN bus. In: 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE (2018)
8. Malik, S., Sun, W.: Analysis and simulation of cyber attacks against connected and autonomous vehicles. In: 2020 International Conference on Connected and Autonomous Driving (MetroCAD). IEEE (2020)
9. Koh, Y., Kim, S., Kim, Y., Oh, I., Yim, K.: Efficient CAN dataset collection method for accurate security threat analysis on vehicle internal network. In: Barolli, L. (ed.) *IMIS 2022. LNCS*, vol. 496, pp. 97–107. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08819-3_10
10. https://documentation.help/PCAN-Basic/PCAN-Basic_Documentation.html



Performance Analysis of HARQ in 2-step RACH Procedure Using Markov Chain Model

Byungchan Kim and Hyunhee Park^(✉)

Myongji University, Yongin-si, Gyeonggi-do, South Korea
{bckim, hhpark}@mj.ac.kr

Abstract. A bit error may occur when data is transmitted in the 5G/6G cellular system. Hybrid Automatic Repeat and reQuest (HARQ) for an error correction and retransmission method is being applied as an error control method to solve this problem. Bit errors can also occur during the MsgA transmission process for the initial connection of the 2-step RACH procedure, a random access method announced in Release 16 of the 3GPP. However, HARQ is not currently applied in the MsgA transmission process. In this paper, we propose a method that applies HARQ to the MsgA transmission process to improve the initial connection of the 2-step RACH procedure. We analyze the transmission process using analytical Markov Chain Model. For performance analysis, the performance of the HARQ-applied method is compared to the performance of the existing transmission method. As a result of comparative analysis, in the case of 256QAM, the HARQ-applied method at saturation of -6 dB increased by about 61% compared to the existing method.

Keywords: 5G/6G · HARQ · Markov chain model · 2-step RACH procedure

1 Introduction

In wireless communication, a bit error may occur during a transmission process. To solve this problem, an error control method is used during the transmission process. Among error correction methods, Forward Error Correction (FEC) of the Physical Layer directly corrects errors in bit errors [1]. Among retransmission methods, Automatic Repeat and reQuest (ARQ) of the Medium Access Control (MAC) layer requests retransmission of an error packet in the event of a bit error [2]. In particular, the hybrid automatic repeat and request (HARQ) method, which uses a combination of two representative methods of error control together, is a method that has been in the spotlight in wireless communication as a method that can effectively improve accuracy and reliability [3].

With the rapid development of cellular systems such as 5G/6G, enhanced mobile broadband (eMBB), ultra-reliable and low latency communication (URLLC), human-central services (HCS), and mobile broadband reliable low latency communication (MBRLLC) are required service categories [4, 5]. In particular, the use of multiple devices, such as Massive Machine Type Communications (mMTC), has resulted in a large amount of data traffic. In 2030, 97Billion of devices are expected to be used,

and the volume of mobile traffic worldwide is expected to increase 670 times in 2030 compared to 2010 [6]. With the high volume of traffic, efficient Random Access Procedure has become more important as a way to improve the connectivity of wireless communication to multiple devices [7]. The Release 16 standard of 3GPP describes the two-step RACH procedure along with the existing “4-step RACH Procedure” random access method. This ensures faster performance and quality by reducing transmission delay than the traditional random access method [8]. The MsgA transfer process for the initial connection of the 2-step RACH procedure based on wireless communication can also be sufficiently error-prone. Therefore, an error control method is also required for the 2-step RACH procedure. The accuracy and reliability of the initial connection of the Random Access procedure can be ensured when HARQ, which is capable of both error correction and retransmission, is applied during the Error control method. However, in Release 16, published by 3GPP, HARQ is not specified in the MsgA transmission process for Physical Uplink Shared Channel (PUSCH) [9].

In this paper, we propose to improve the initial connection by applying HARQ to the MsgA transmission process of the 2-step RACH Procedure. In addition, to analyze the performance of the HARQ-applied transmission method, we categorize the HARQ states that occur in the two-stage RACH procedure into 5 categories and create a Markov chain model with state transition probabilities. Numerical performance analysis and simulation were performed using this Markov chain model. For performance analysis, we compare 2-step RACH Procedure with and without HARQ.

2 2-step RACH Procedure

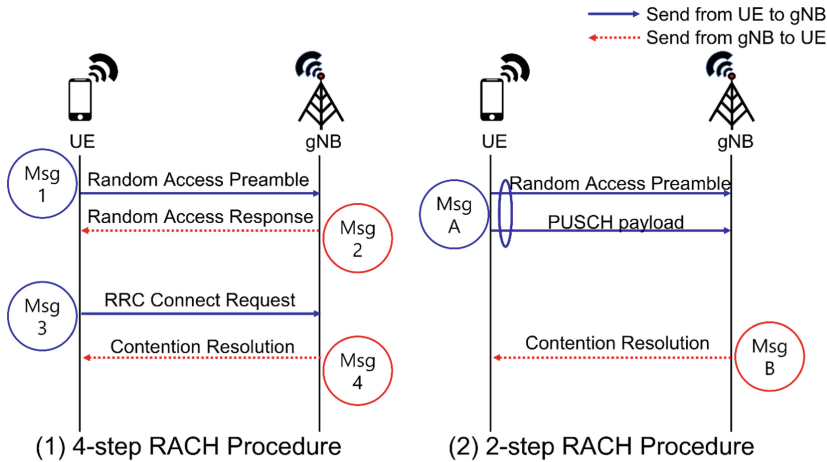


Fig. 1. RACH procedure.

Random access is a method for establishing a connection between a terminal and a base station, and is classified into Contention-Based Random Access and Contention-Free Random Access [10]. Contention-based random access is a method in which multiple (a plurality of) terminals compete with each other to use one channel, and representative examples include ALOHA, CSMA/CD, CSMA/CA methods, etc. Contention-free random access is a method in which there is no channel competition between each terminal, and representative examples include token bus and round robin methods. We deal with contention-based random access.

Recently, many studies have been conducted to reduce the delay time of the random access procedure [11]. In particular, the 2-step RACH procedure published in 3GPP is a representative example [12]. In the case of the 4-step RACH Procedure, which uses the existing four-hand shaking method, the Random Access is performed as a process of sending preamble and confirming connection, as shown in Fig. 1 (1). For the 2-step RACH procedure, we propose a method for transmitting Msg1 and Msg3 of the four-step RACH procedure as MsgA, as shown in Fig. 1 (2). This method is a method that can reduce latency and reduce signaling. However, since the 2-step RACH procedure is also wireless communication, we find that errors can occur when transmitting MsgA, and study a method to solve this problem.

3 Hybrid Automatic Repeat and Request

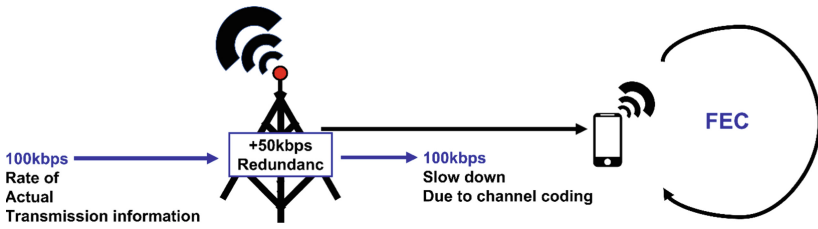


Fig. 2. FEC operation process.

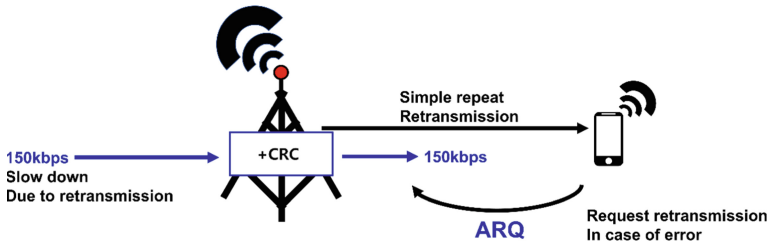


Fig. 3. ARQ operation process.

Methods for performing error control are largely classified as error correction and retransmission. As a representative example of error correction, the FEC of the physical

layer described in Fig. 2, transmits a message by adding surplus bits at the transmitting side. When an error occurs in the bit, it restore the error by itself. As a representative example of retransmission, the ARQ of the MAC layer described in Fig. 3, refers to a method of retransmitting data from a point where an error occurs during NACK or Timeout.

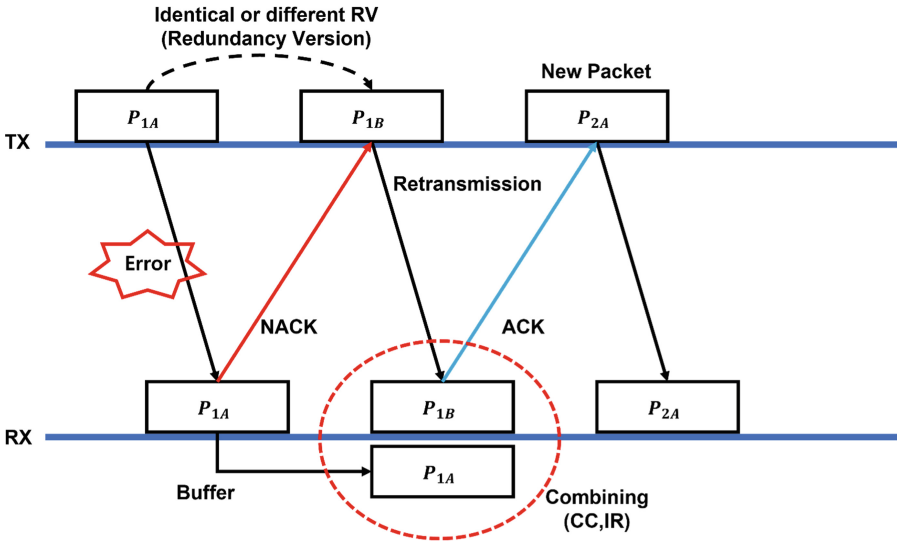


Fig. 4. HARQ operation process.

As you can see in Fig. 4, HARQ is a combination of FEC and ARQ, and when an error occurs on the receiving side, HARQ stores the error-occurring packet in the buffer and requests packet retransmission [13]. The packet stored in the buffer and the retransmitted packet are combined to restore the original packet. Types of HARQ are classified according to how packets are retransmitted and combined. They are the “Incremental Redundancy” (IR) method and the “Chase Combining” (CC) method. The HARQ method enables efficient error control by performing both error correction and retransmission.

4 Analytical Markov Chain Model

We found the problem that errors can also occur in the MsgA transmission process of 2-step RACH Procedure. However, HARQ is not currently applied in the current MsgA transmission process. In this paper, we propose a method for applying HARQ to MsgA transmission process to improve the initial connection of the 2-step RACH procedure.

It is assumed that all equations are derived from an Additive White Gaussian Noise (AWGN) environment. Since the bit error probability depends on the channel environment, it is necessary to obtain a signal noise ratio (SNR) value. The SNR value can be

obtained as shown in Eq. (1) from [14].

$$\gamma_b := \frac{E_b}{N_0} = \frac{A^2}{N_0} = \frac{d_{min}^2}{4}. \quad (1)$$

SNR is an average signal power versus average noise power and is a value indicating a degree to which noise affects a signal. E_b represents bit energy, N_0 represents spectrum density of noise power, and power per 1 Hz. For E_b , it can be represented by the band of the transmission signal, which is the bandwidth given by $x \in \{-A, A\}$, which can be represented as d_{min} in the constellation according to the modulation scheme. $\frac{E_b}{N_0}$ is a parameter indicating energy per bit required for noise power per unit bandwidth and is expressed as γ_b . It is also called SNR per bit [14].

Accordingly, the bit error probability be expressed as a Q-function as in Eqs. (2.1) and (2.2) from [15].

$$P_e = P\{n > A\} = \int_A^\infty \frac{1}{\sqrt{2\pi\sigma^2/2}} e^{-\frac{x^2}{2\sigma^2/2}} = Q\left(\sqrt{\frac{d^2}{2N_0}}\right) = Q(\sqrt{2\gamma_b}). \quad (2.1)$$

$$P_e \cong \frac{4}{\log_2 M} Q\left(\sqrt{\frac{3\gamma_b \log_2 M}{M-1}}\right). \quad (2.2)$$

The bit error probability may be expressed as in (2.1) and (2.2) according to modulation. Equations (2.1) and (2.2) refer to bit error probabilities in the QPSK environment and the M-QAM environment, respectively.

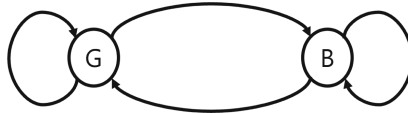


Fig. 5. Gilbert-Elliott model.

Through the bit error probability according to the SNR value, the channel can be classified into two states: Good Channel and Bad Channel. We can calculate the state transition probability through the 2-state Markov Chain Model, known as the Gilbert-Elliott Model in Fig. 5, and obtain the Bit Error Rate (BER) as Eq. (3) [16].

$$BER = \varepsilon P_e(1) + (1 - \varepsilon) P_e(2). \quad (3)$$

ε means the probability that the channel is in the Bad state in the normal state, and $P_e(*)$ means the probability that an error occurs in the state * [17].

We intend to apply HARQ to the transmission process of MsgA in the 2-step RACH Procedure. The transmission of MsgA does not only send Preamble, but also PUSCH Payload, so throughput calculations per packet are required. The equation related to the packet error ratio (PER) and the BER can be obtained through Eq. (4), and L means the length of the packet [18].

$$PER = 1 - (1 - BER)^L. \quad (4)$$

The throughput in packets may be calculated as shown in Eq. (5) from [18]. The values for obtaining throughput vary according to each modulation.

$$Throughput = (1 - PER) \times Symbol_{rate} \times bit_per_symbol \times Code_rate. \quad (5)$$

Table 1. State table of HARQ

State	Definition
T_0	New Packet Transfer Attempt State After Previous Transfer Success
T_f	New Packet Transfer Attempt Status After Previous Transfer Failure
E	Packet Error State
D	Delay State after Packet Error until retransmission state
R	Retransmission State

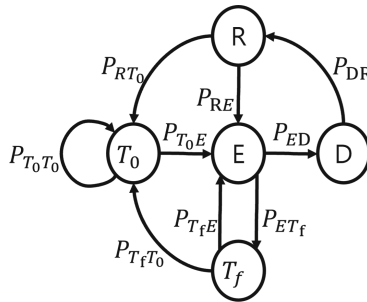


Fig. 6. Markov chain model for HARQ in 2-step RACH procedure.

Table 1 classifies the states occurring in HARQ into 5 categories [19]. In this paper, the probabilistic analysis is conducted by creating a Markov Chain Model by calculating the state transition probability based on the state of HARQ in Table 1. This is shown in Fig. 6.

$$Matrix = \begin{bmatrix} P_{T_0T_0} & 0 & P_{T_0E} & 0 & 0 \\ P_{T_fT_0} & 0 & P_{T_0E} & 0 & 0 \\ 0 & P_{T_fE} & 0 & P_{ED} & 0 \\ 0 & 0 & 0 & 0 & P_{DR} \\ P_{RT_0} & 0 & P_{RE} & 0 & 0 \end{bmatrix}. \quad (6)$$

Equation (6) shows the transmission probability for each state in the Markov Chain Model of HARQ analytical Markov Chain Model in Fig. 6 as Matrix. For example, P_{DR}

refers to the probability from the delay state D to the retransmission state R after the packet error.

$$\pi = [\pi_{T_0} \ \pi_{T_f} \ \pi_D \ \pi_C \ \pi_R] \quad (7)$$

In Eq. (7), π_X represents the probability of being in state X in a steady state. Therefore, steady state probability vectors are denoted as $\pi = \sum \pi$ [20].

$$E\{g\} = \pi_{T_0} * 1^T. \quad (8)$$

$$E\{d\} = (\pi_{T_0} * 1^T) + (r * N_{T_0} * \pi_{T_f} * 1^T) + (\pi_C * 1^T) + (r * \pi_D * [1 \dots N_0]^T). \quad (9)$$

Equations (8) and (9) respectively represent (mean) a successful average packet gain and an average delay in a steady state. r represents the ratio of the counter step duration to the duration of the competition period, channel access is performed to the node with a smaller counter value, and the initial counter value randomly selects the size N_0 of the window. When a collision occurs, N_0 increases to $2N_0$, and the window size of N_{T_0} increases to the predetermined size.

$$\text{Average Throughput} = \frac{E\{\text{total number of successful packets}\}}{E\{\text{total delay}\}/\text{slot duration}} = \frac{E\{g\}}{E\{d\}}. \quad (10)$$

Equation (10) for average throughput can be obtained through Eqs. (8) and (9).

5 Simulation

The performance evaluation was conducted when HARQ was applied in the MsgA transmission process of 2-step RACH Procedure using the Markov Chain Model proposed in the section above. To this end, performance evaluation was conducted with Matlab using the following simulation parameters in Table 2.

Table 2. Simulation parameters

Description	Value
Modulation	QPSK, 16QAM, 64QAM, 256QAM
Channel Coding	LDPC
Code rate	1/4
HARQ Process	16
Initial Backoff Counter	8
Number of symbols	14
Eb/N0	0–25
Packet length	32

When sending MsgA, PUSCH payload and PRACH preamble are sent together. Therefore, among the standards of 3GPP, the defined modulation and channel coding parts were applied when Physical Uplink Shared Channel (PUSCH) was used in a 5G channel environment [21]. HARQ Process also substituted the value of 3GPP when using a given number of HARQ Processes in PUSCH [22]. Set the initial Backoff Window to 8 and set r to 0.05. The number of symbols is set to 14 [23]. Set the packet length to 32 bytes [24]. It is assumed that error correction is performed by 1/4 during bit error and retransmission is performed during bit error.

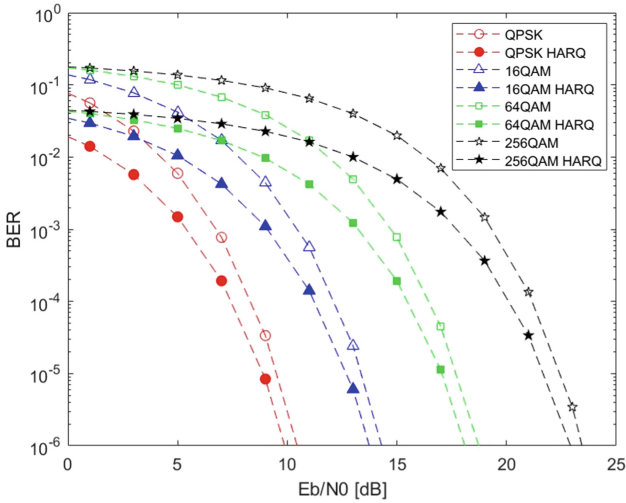


Fig. 7. E_b/N_0 [dB] versus BER.

Figure 7 compares the BER of the existing 2-step RACH Procedure and the BER of the method of HARQ-applied method for each modulation. The BER curve of the existing 2-step RACH Procedure is represented by a dashed line with empty symbol, and the method to which HARQ is applied is indicated by coloring the interior. Compared with each modulation, if the channel performance is good, that is, if E_b/N_0 is high, all of them draw a downward curve and BER is reduced. Furthermore, the HARQ-applied method shows better performance compared to the existing method.

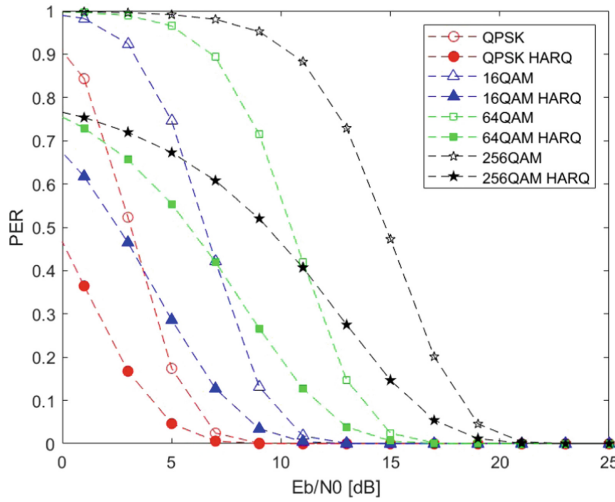


Fig. 8. E_b/N_0 [dB] versus PER.

Figure 8 was prepared by calculating the value of PER formed according to BER for each modulation. As in Fig. 7, it can be confirmed that the error ratio decreases when HARQ is applied.

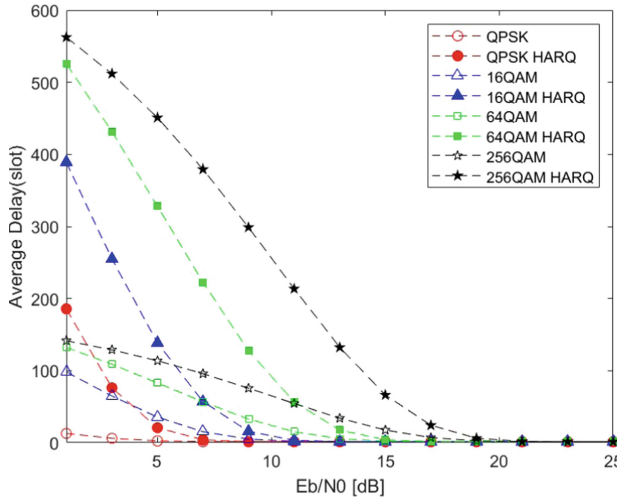


Fig. 9. E_b/N_0 [dB] versus average delay.

Figure 9 is obtained by calculating the average delay in a normal state through the probability, and it can be verified that the part to which HARQ is applied takes more time than the existing method because the retransmission process continues.

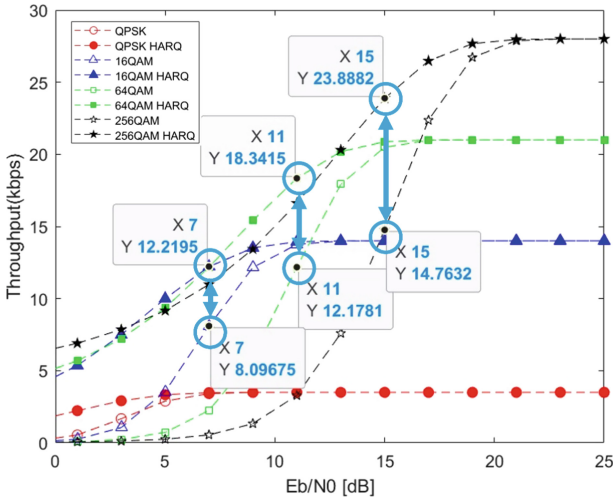


Fig. 10. E_b/N_0 [dB] versus throughput.

Table 3. Throughput increase rate in HARQ

Modulation	Saturation E_b/N_0	Saturation $E_b/N_0 - 6$ dB	Throughput increase rate in HARQ (%)
16QAM	13	7	150.4431
64QAM	17	11	150.6105
256QAM	21	15	161.8091

From Fig. 10, it is possible to check the performance when HARQ is applied to the MsgA transmission process of the 2-step RACH Procedure. If E_b/N_0 is good, both the existing 2-step RACH Procedure method and the HARQ-applied method can reach saturation. However, it can be seen from Table 3 that there is a significant difference between the HARQ-applied method and the existing 2-step RACH Procedure method in the saturation state of -6 dB. When QAM is used for modulation, the average delay is large, but the HARQ-applied method confirms that the throughput increases by about 50%. Therefore, when there is a lot of noise, that is, when the channel environment is not good, 2-step RACH procedure is performed using HARQ, which is better than the existing method in the initial connection in all modulations.

6 Conclusion

When data is transmitted in a wireless communication system, errors appear depending on the channel environment. There are various error control methods to solve this problem. Transmitting MsgA in the 2-step RACH Procedure is also a wireless communication process, which can cause problems. In this paper, We improve and analyze the

initial connection by applying HARQ that was not in the MsgA transfer process of the second stage RACH procedure. In addition, we create a Markov chain model and perform performance evaluation compared to the existing 2-step RACH Procedure method. In Particular, in all QAMs, the throughput of the HARQ-applied method increased by about 50% compared to the existing 2-step RACH Procedure without HARQ at the saturation state of -6 dB, confirming that the performance was excellent.

However, the analytical Markov Chain Model does not specify the formula for Error Correction. In the future, we plan to conduct research by replacing the exact error correction process.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded and Institute for Information & communications Technology Planning & Evaluation(IITP) by the Korea government (MSIT)(No. 2022R1A2C2005705, AI-MAC Protocol on Distributed Machine Learning for Intelligent Flying Base Station, No. 2021-0-00368, Development of the 6G Service Targeted AI/ML-based autonomous-Regulating Medium Access Control (6G STAR-MAC)).

References

1. Rowitch, D.N., Milstein, L.B.: On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes. *IEEE Trans. Commun.* **48**(6), 948–959 (2000)
2. Vaze, R.: Throughput-delay-reliability tradeoff with ARQ in wireless ad hoc networks. *IEEE Trans. Wireless Commun.* **10**(7), 2142–2149 (2011)
3. Lott, C., Milenkovic, O., Soljanin, E.: Hybrid ARQ: theory, state of the art and future directions. In: 2007 IEEE Information Theory Workshop on Information Theory for Wireless Networks, pp. 1–5. IEEE (2007)
4. Popovski, P., et al.: 5G wireless network slicing for eMBB, URLLC, and mMTC: a communication-theoretic view. *IEEE Access* **6**, 55765–55779 (2018)
5. Saad, W., Bennis, M., Chen, M.: A vision of 6G wireless systems: applications, trends, technologies, and open research problems. *IEEE Netw.* **34**(3), 134–142 (2019)
6. Chowdhury, M.Z., et al.: 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions. *IEEE Open J. Commun. Soc.* **1**, 957–975 (2020)
7. Laya, A., Alonso, L., Alonso-Zarate, J.: Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives. *IEEE Commun. Surv. Tutor.* **16**(1), 4–16 (2013)
8. 3GPP TR21.916. Release 16 Description; Summary of Rel-16 Work Items, January 2020
9. 3GPP TSG RAN #88e RP-200622. NR_2step_RACH, 29 June–3 July 2020
10. Leyva-Mayorga, I., et al.: Random Access for Machine-Type Communications. Wiley 5G Ref: The Essential 5G Reference Online, pp. 1–21 (2019)
11. Cheng, J.-P., Lee, C., Lin, T.-M.: Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks. In: 2011 IEEE GLOBECOM Workshops (GC Wkshps), pp. 368–372. IEEE (2011)
12. 3GPP TS38.321. NR; Medium Access Control (MAC) protocol specification, May 2022
13. Ahmed, A., et al.: Hybrid automatic repeat request (HARQ) in wireless communications systems and standards: a contemporary survey. *IEEE Commun. Surv. Tutor.* **23**(4), 2711–2752 (2021)

14. Tse, D., Viswanath, P.: *Fundamentals of Wireless Communication*. Cambridge University Press, Cambridge (2005)
15. Meghdadi, V.: BER calculation. *Wireless Communications* (2008)
16. Gilbert, E.N.: Capacity of a burst-noise channel. *Bell Syst. Tech. J.* **39**(5), 1253–1265 (1960)
17. Kang, J.H., Stark, W.E., Hero, A.O.: Turbo codes for fading and burst channels. In: *IEEE Mini Conference at Globecom*, pp. 40–45 (1998)
18. Ito, K.: Performance evaluation and improvement of PER and throughput in galvanic-coupling intra-body communication systems. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3742–3745. IEEE (2018)
19. Kim, B., Park, H.: Markov chain model of HARQ based on retransmission for 2-step RACH procedure. In: *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, pp. 203–204 (2022)
20. Tutgun, R., Aktas, E.: A Markovian analysis of cooperative ARQ with random access. *Wirel. Pers. Commun.* **123**(4), 3201–3211 (2022)
21. 3GPP TR21.915. Release 15 Description; Summary of Rel-15 Work Items, October 2019
22. 3GPP TS38.214. NR; Physical layer procedures for data, July 2022
23. 3GPP TSG RAN meeting#88e RP-200622. NR_2step_RACH, 29 June–3 July 3 2020
24. Segura, D., et al.: 5G numerologies assessment for URLLC in industrial communications. *Sensors* **21**(7), 2489 (2021)



A Comparison Study of FC-RDVM with LDVM Router Replacement Methods by WMN-PSOHC Simulation System Considering Weibull Distribution of Mesh Clients

Shinji Sakamoto¹(✉), Admir Barolli², Yi Liu³, Elis Kulla⁴, Leonard Barolli⁵,
and Makoto Takizawa⁶

¹ Department of Information and Computer Science, Kanazawa Institute
of Technology, 7-1 Ohgigaoka, Nonoichi, Ishikawa 921-8501, Japan
shinji.sakamoto@ieee.org

² Department of Information Technology, Aleksander Moisiu University of Durres,
L.I, Rruga e Currilave, Durres, Albania
admirbarolli@uamd.edu.al

³ Department of Computer Science, National Institute of Technology, Oita College,
1666, Maki, Oita 870-0152, Japan
y-liu@oita-ct.ac.jp

⁴ Department of System Management, Fukuoka Institute of Technology,
3-30-1 Wajiro-Higashi, Higashi-ku, Fukuoka 811-0295, Japan
kulla@fit.ac.jp

⁵ Department of Information and Communication Engineering, Fukuoka Institute
of Technology, 3-30-1 Wajiro-Higashi, Higashi-ku, Fukuoka 811-0295, Japan
barolli@fit.ac.jp

⁶ Department of Advanced Sciences, Faculty of Science and Engineering,
Hosei University, Kajino-Machi, Koganei-shi, Tokyo 184-8584, Japan
makoto.takizawa@computer.org

Abstract. Wireless Mesh Networks (WMNs) have many advantages such as easy maintenance, low up-front cost, high robustness. However, WMNs have some problems such as node placement problem, security, transmission power and so on. In this work, we deal with node placement problem. In our previous work, we implemented a hybrid simulation system based on Particle Swarm Optimization (PSO) and Hill Climbing (HC) called WMN-PSOHC for solving the node placement problem in WMNs. In this paper, we present the implementation of Fast Convergence RDVM (FC-RDVM) in WMN-PSOHC. Then, we evaluate the performance of WMNs by using WMN-PSOHC hybrid simulation system. We compare the performance of FC-RDVM with LDVM considering Weibull distribution of mesh clients. Simulation results show that FC-RDVM performs better than LDVM

1 Introduction

The Wireless Mesh Networks (WMNs) have many advantages compared with conventional Wireless Local Area Networks (WLANs) such as low-up front cost,

easy deployment and high robustness [1,6]. Also, they provide better services than conventional WLANs. The WMNs can be applied for medical, transport and surveillance applications in urban area, metropolitan area, neighbouring communities and municipal area networks [2,5,7]. Nodes of WMNs can be categorized into four kinds of nodes according to their roles: Mesh Point (MP), Mesh Access Point (MAP), Mesh Portal (MPP) collocated with MP and Stations (STA). The STA is a conventional WLAN's client node. The MP does not provide the Internet connection to STA, but it supports Peer Link Management Protocol (PLMP). The MAP is MP providing the Internet connection to STA. The MPP is a node having a gateway feature in a WMN, connected by cable.

However, there are some problems to be solved in WMNs. One of the critical issues of WMNs is the achievement of network connectivity and user coverage, which is closely related to the family of node placement problems. The node placement problem has been investigated in the optimization field and is known as the class of NP-hard.

We already implemented a Particle Swarm Optimization (PSO) based simulation system, called WMN-PSO [10]. Also, we implemented a simulation system based on Hill Climbing (HC) for solving node placement problem in WMNs, called WMN-HC [9].

In our previous work, we presented a hybrid intelligent simulation system based on PSO and HC [11]. We called this system WMN-PSOHC. We also implemented Linearly Decreasing Vmax Method (LDVM) replacement method in WMN-PSOHC [12]. In this paper, we compare the performance of FC-RDVM with LDVM. Simulation results show that FC-RDVM has better performance than LDVM.

The rest of the paper is organized as follows. In Sect. 2, we present intelligent algorithms. We present the implemented hybrid simulation system and FC-RDVM in Sect. 3. The simulation results are given in Sect. 4. Finally, we give conclusions and future work in Sect. 5.

2 Intelligent Algorithms

2.1 Particle Swarm Optimization

In Particle Swarm Optimization (PSO) algorithm, a number of simple entities (the particles) are placed in the search space of some problem or function and each evaluates the objective function at its current location. The objective function is often minimized and the exploration of the search space is not through evolution [8]. However, following a widespread practice of borrowing from the evolutionary computation field, in this work, we consider the bi-objective function and fitness function interchangeably. Each particle then determines its movement through the search space by combining some aspect of the history of its own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have been moved. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function.

Algorithm 1. Pseudo code of WMN-PSOHC.

```

/* Generate the initial solutions and parameters */
Computation maxtime:=  $T_{max}$ ,  $k := 1$ ;
Number of particle-patterns:=  $m$ ,  $2 \leq m \in \mathbf{N}^1$ ;
Particle-patterns initial solution:=  $\mathbf{P}_i^0$ ;
Global initial solution:=  $\mathbf{G}^0$ ;
Particle-patterns initial position:=  $\mathbf{x}_{ij}^0$ ;
Particles initial velocity:=  $\mathbf{v}_{ij}^0$ ;
PSO parameter:=  $\omega$ ,  $0 < \omega \in \mathbf{R}^1$ ;
PSO parameter:=  $C_1$ ,  $0 < C_1 \in \mathbf{R}^1$ ;
PSO parameter:=  $C_2$ ,  $0 < C_2 \in \mathbf{R}^1$ ;
/* Start PSO-HC */
Evaluate( $\mathbf{G}^0$ ,  $\mathbf{P}^0$ );
while  $k \leq T_{max}$  do
  /* Update velocities and positions */
   $\mathbf{v}_{ij}^{k+1} = \omega \cdot \mathbf{v}_{ij}^k$ 
     $+ C_1 \cdot \text{rand}() \cdot (\text{best}(\mathbf{P}_{ij}^k) - \mathbf{x}_{ij}^k)$ 
     $+ C_2 \cdot \text{rand}() \cdot (\text{best}(\mathbf{G}^k) - \mathbf{x}_{ij}^k)$ ;
   $\mathbf{x}_{ij}^{k+1} = \mathbf{x}_{ij}^k + \mathbf{v}_{ij}^{k+1}$ ;
  /* if fitness value is increased, a new solution will be accepted. */
  if Evaluate( $\mathbf{G}^{(k+1)}$ ,  $\mathbf{P}^{(k+1)}$ )  $\geq$  Evaluate( $\mathbf{G}^{(k)}$ ,  $\mathbf{P}^{(k)}$ ) then
    Update_Solutions( $\mathbf{G}^k$ ,  $\mathbf{P}^k$ );
    Evaluate( $\mathbf{G}^{(k+1)}$ ,  $\mathbf{P}^{(k+1)}$ );
  else
    ReUpdate_Solutions( $\mathbf{G}^{k+1}$ ,  $\mathbf{P}^{k+1}$ );
  end if
   $k = k + 1$ ;
end while
Update_Solutions( $\mathbf{G}^k$ ,  $\mathbf{P}^k$ );
return Best found pattern of particles as solution;

```

Each individual in the particle swarm is composed of three \mathcal{D} -dimensional vectors, where \mathcal{D} is the dimensionality of the search space. These are the current position \mathbf{x}_i , the previous best position \mathbf{p}_i and the velocity \mathbf{v}_i .

The particle swarm is more than just a collection of particles. A particle by itself has almost no power to solve any problem. The progress occurs only when the particles interact. Problem solving is a population-wide phenomenon, emerging from the individual behaviors of the particles through their interactions. In any case, populations are organized according to some sort of communication structure or topology, often thought of as a social network. The topology typically consists of bidirectional edges connecting pairs of particles, so that if j is in i 's neighborhood, i is also in j 's. Each particle communicates with some other particles and is affected by the best point found by any member of its topological neighborhood. This is just the vector \mathbf{p}_i for that best neighbor, which we will denote with \mathbf{p}_g . The potential kinds of population "social networks" are hugely varied, but in practice certain types have been used more frequently.

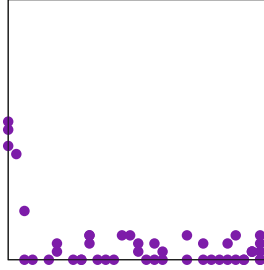


Fig. 1. Weibull distribution of mesh clients.

In the PSO process, the velocity of each particle is iteratively adjusted so that the particle stochastically oscillates around \mathbf{p}_i and \mathbf{p}_g locations.

2.2 Hill Climbing

Hill Climbing (HC) algorithm is a heuristic algorithm. The idea of HC is simple. In HC, the solution s' is accepted as the new current solution if $\delta \leq 0$ holds, where $\delta = f(s') - f(s)$. Here, the function f is called the fitness function. The fitness function gives points to a solution so that the system can evaluate the next solution s' and the current solution s .

The most important factor in HC is to define effectively the neighbor solution. The definition of the neighbor solution affects HC performance directly. In our WMN-PSOHC system, we use the next step of particle-pattern positions as the neighbor solutions for the HC part.

3 WMN-PSOHC Hybrid Simulation System and FC-RDVM

We show the pseudo code of WMN-PSOHC in Algorithm 1. In following, we present the initialization, particle-pattern, fitness function and router replacement methods.

Initialization

Our proposed system starts by generating an initial solution randomly, by *ad hoc* methods [16]. We decide the velocity of particles by a random process considering the area size. For instance, when the area size is $W \times H$, the velocity is decided randomly from $-\sqrt{W^2 + H^2}$ to $\sqrt{W^2 + H^2}$. Our system can generate many client distributions. In this paper, we consider Weibull distribution of mesh clients as shown in Fig. 1.

Particle-Pattern

A particle is a mesh router. A fitness value of a particle-pattern is computed by combination of mesh routers and mesh clients positions. In other words,

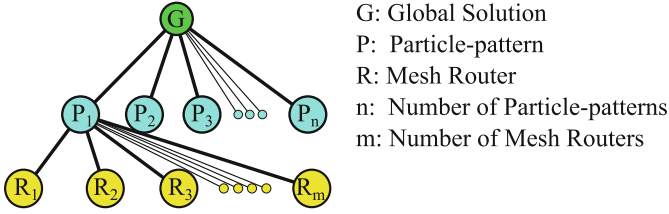


Fig. 2. Relationship among global solution, particle-patterns and mesh routers.

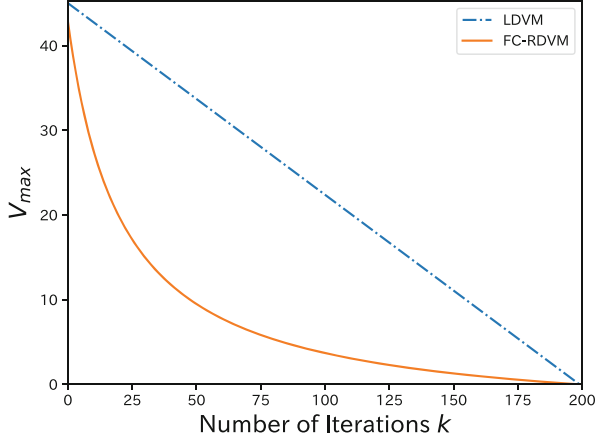


Fig. 3. The difference of V_{max} between LDVM and FC-RDVM.

each particle-pattern is a solution as shown in Fig. 2. Therefore, the number of particle-patterns is the number of solutions.

Fitness Function

One of most important thing is to decide the determination of an appropriate objective function and its encoding. In our case, each particle-pattern has an own fitness value and compares other particle-patterns fitness value in order to share information of global solution. The fitness function follows a hierarchical approach in which the main objective is to maximize the SGC in WMN. Thus, we use α and β weight-coefficients for the fitness function and the fitness function of this scenario is defined as:

$$\text{Fitness} = \alpha \times \text{SGC}(\mathbf{x}_{ij}, \mathbf{y}_{ij}) + \beta \times \text{NCMC}(\mathbf{x}_{ij}, \mathbf{y}_{ij}).$$

Router Replacement Methods

A mesh router has x, y positions and velocity. Mesh routers are moved based on velocities. There are many router replacement methods in PSO field [4, 13–15]. In this paper, we compare two replacement methods: Linearly Decreasing Vmax Method (LDVM) and Fast Convergence Rational Decrement of Vmax Method (FC-RDVM).

Table 1. Parameter settings.

Parameters	Values
Clients distribution	Weibull distribution
Area size	32×32
Number of mesh routers	16
Number of mesh clients	48
Total iterations	800
Iteration per phase	4
Number of particle-patterns	9
Radius of a mesh router	From 2.0 to 3.0
Fitness function weight-coefficients (α, β)	0.7, 0.3
Curvature parameter (γ)	10.0
Replacement methods	LDVM, FC-RDVM

In LDVM, PSO parameters are set to unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). A value of V_{max} which is maximum velocity of particles is considered. With increasing of iteration of computations, the V_{max} is kept decreasing linearly [3, 13]. The V_{max} is defined as shown in Eq. (1).

$$V_{max}(k) = \sqrt{W^2 + H^2} \times \frac{T - k}{T} \quad (1)$$

where W and H are the width and the height of the considered area, respectively. Also, T and k are the total number of iterations and a current number of iteration, respectively. The k is a variable varying from 1 to T , which is increased by increasing the iterations.

In FC-RDVM, the V_{max} decreases with the increasing of iterations as shown in Eq. (2).

$$V_{max}(k) = \sqrt{W^2 + H^2} \times \frac{T - k}{T + \gamma k} \quad (2)$$

where γ is a curvature parameter. When the γ is larger, the curvature is larger as shown in Fig. 3. Other parameters are the same with LDVM.

4 Simulation Results

In this section, we show simulation results using WMN-PSOHC hybrid intelligent system. In this work, we consider Weibull distribution of mesh clients. We consider the number of particle-patterns 9. We conducted simulations 100 times in order to avoid the effect of randomness and create a general view of results. The total number of iterations is considered 800 and the iterations per phase is considered 4. We show the parameter setting for WMN-PSOHC in Table 1.

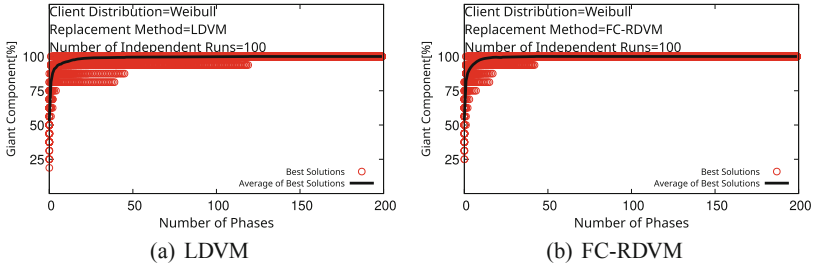


Fig. 4. Simulation results of WMN-PSOHC for SGC.

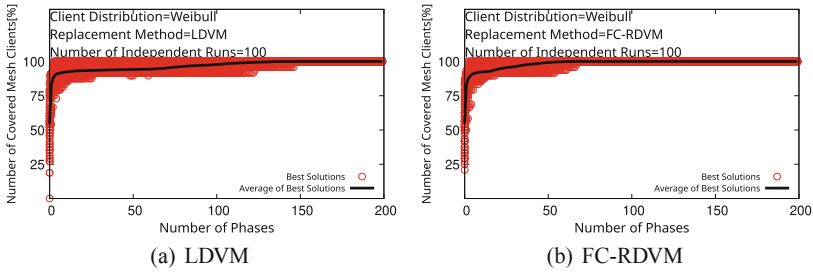


Fig. 5. Simulation results of WMN-PSOHC for NCMC.

We show the simulation results in Fig. 4 and Fig. 5. Considering SGC parameter, the LDVM convergence is slower than FC-RDVM. For NCMC, the FC-RDVM converges faster than LDVM. When we use FC-RDVM, the performance converges to 100% before 100 phases. On the other hand, in the case of LDVM, the performance converges to 100% for 150 phases.

5 Conclusions

In this work, we evaluated the performance of WMNs by using WMN-PSOHC hybrid simulation system. We also proposed and implemented FC-RDVM router replacement method. Then, we compared the performance of FC-RDVM with LDVM considering Weibull distribution of mesh clients. Simulation results show that FC-RDVM performs better than LDVM.

In our future work, we would like to evaluate the performance of the proposed system for different parameters and scenarios.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. *Comput. Netw.* **47**(4), 445–487 (2005)
2. Amaldi, E., Capone, A., Cesana, M., Filippini, I., Malucelli, F.: Optimization models and methods for planning wireless mesh networks. *Comput. Netw.* **52**(11), 2159–2171 (2008)

3. Barolli, A., Sakamoto, S., Ohara, S., Barolli, L., Takizawa, M.: Performance analysis of WMNs by WMN-PSOHC-DGA simulation system considering random inertia weight and linearly decreasing Vmax router replacement methods. In: Barolli, L., Hussain, F.K., Ikeda, M. (eds.) CISIS 2019. AISC, vol. 993, pp. 13–21. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-22354-0_2
4. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evol. Comput.* **6**(1), 58–73 (2002)
5. Franklin, A.A., Murthy, C.S.R.: Node placement algorithm for deployment of two-tier wireless mesh networks. In: *Proceedings of Global Telecommunications Conference*, pp. 4823–4827 (2007)
6. Islam, M.M., Funabiki, N., Sudibyo, R.W., Munene, K.I., Kao, W.C.: A dynamic access-point transmission power minimization method using PI feedback control in elastic WLAN system for IoT applications. *Internet Things* **8**(100), 089 (2019)
7. Muthaiah, S.N., Rosenberg, C.P.: Single gateway placement in wireless mesh networks. In: *Proceedings of 8th International IEEE Symposium on Computer Networks*, pp. 4754–4759 (2008)
8. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. *Swarm Intell.* **1**(1), 33–57 (2007)
9. Sakamoto, S., Lala, A., Oda, T., Kolici, V., Barolli, L., Xhafa, F.: Analysis of WMN-HC simulation system data using Friedman test. In: *The Ninth International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2015)*, pp. 254–259. IEEE (2015)
10. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation and evaluation of a simulation system based on particle swarm optimisation for node placement problem in wireless mesh networks. *Int. J. Commun. Netw. Distrib. Syst.* **17**(1), 1–13 (2016)
11. Sakamoto, S., Ozera, K., Ikeda, M., Barolli, L.: Implementation of intelligent hybrid systems for node placement problem in WMNs considering particle swarm optimization, hill climbing and simulated annealing. *Mob. Netw. Appl.* **23**(1), 27–33 (2017). <https://doi.org/10.1007/s11036-017-0897-7>
12. Sakamoto, S., Barolli, L., Okamoto, S.: A comparison study of linearly decreasing inertia weight method and rational decrement of Vmax method for WMNs using WMN-PSOHC intelligent system considering normal distribution of mesh clients. In: Barolli, L., Natwichai, J., Enokido, T. (eds.) *EIDWT 2021. LNDECT*, vol. 65, pp. 104–113. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70639-5_10
13. Schutte, J.F., Groenwold, A.A.: A study of global optimization using particle swarms. *J. Global Optim.* **31**(1), 93–108 (2005)
14. Shi, Y.: Particle swarm optimization. *IEEE Connections* **2**(1), 8–13 (2004)
15. Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E. (eds.) *EP 1998. LNCS*, vol. 1447, pp. 591–600. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0040810>
16. Xhafa, F., Sanchez, C., Barolli, L.: Ad hoc and neighborhood search methods for placement of mesh routers in wireless mesh networks. In: *Proceedings of 29th IEEE International Conference on Distributed Computing Systems Workshops (ICDCS-2009)*, pp. 400–405 (2009)



A Fuzzy-Based System for Estimation of Landslide Disasters Risk Considering Digital Elevation Model

Kei Tabuchi¹, Kyohei Toyoshima², Nobuki Saito², Aoto Hirata², Yuki Nagai², Tetsuya Oda¹(✉), and Leonard Barolli³

¹ Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
t19j048tk@ous.jp, oda@ous.ac.jp

² Graduate School of Engineering, Okayama University of Science (OUS), Okayama, 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
{t22jm24jd,t21jm01md,t21jm02zr,t22jm23rv}@ous.jp

³ Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
barolli@fit.ac.jp

Abstract. Recently, the number of landslide disasters is increased because of heavy rains. For measuring the landslide disasters, it is necessary to consider the characteristics of the mountain topography in addition to rainfalls. Fuzzy inference is a good approach for estimation of disaster risk considering rainfalls and topography parameters. Detecting landslide disasters before they happen requires data collection on the wide area. However, monitoring the entire area of a mountain requires a large number of sensors. In this paper, we present Fuzzy-based system that estimates Landslide Disasters Risk (LDR) considering Digital Elevation Model (DEM). The evaluation results show that the proposed system can estimate LDR according to the rainfall and topography parameter using the real data collected on wide areas by a Wireless Sensor Network (WSN).

1 Introduction

The number of landslide disasters is increased because the amount of rainfalls is increased caused by global warming. Japan is mountainous country and has fragile soils and steep mountains. In addition, Japan has many landslide disasters caused by typhoons and heavy rains [1–3]. Therefore, the estimation and measurement of landslide disasters is very important in order to save the life of the people leaving in close to these areas, immediately. There are some research works that measure the landslide disasters by considering geographic information based on topographic analysis and monitoring by Wireless Sensor Network (WSN) [4–9]. On the other hand, it is hard to predict landslide disasters because of different factors such as mountain geology, topography and amount of rainfalls [10–18].

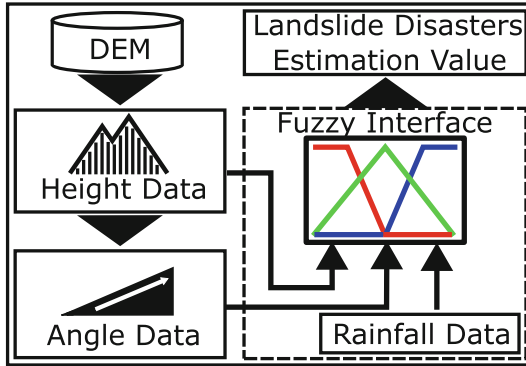


Fig. 1. Proposed system.

Fuzzy inference is a good approach for estimation of disaster risk considering rainfalls and topography parameters. Detecting landslide disasters before they happen requires data collection on the wide area. However, monitoring the entire area of a mountain requires a large number of sensors. Therefore, the number of sensors should be reduced and they should be placed in a way that they can sense the landslides areas. We consider a soil moisture meter for measuring the changes in the moisture content of soil and a tipping bucket rain gauge for measuring the rainfall by $0.5 [mm]$ increments.

In this paper, we present a Fuzzy-based Approach for estimation of Landslide Disasters Risk (LDR) considering Digital Elevation Model (DEM) [19,20]. The evaluation results show that the proposed system can estimate LDR according to the rainfall and topography parameters using the real data collected on wide areas by a Wireless Sensor Network (WSN) [21–24].

The structure of the paper is as follows. In Sect. 2, we present the proposed system. In Sect. 3, we discuss the experimental results. Finally, conclusions and future work are given in Sect. 4.

2 Proposed System

In this section, we describe the proposed system, which estimates LDR and danger areas by landslide disasters. The structure of the proposed system is shown in Fig. 1.

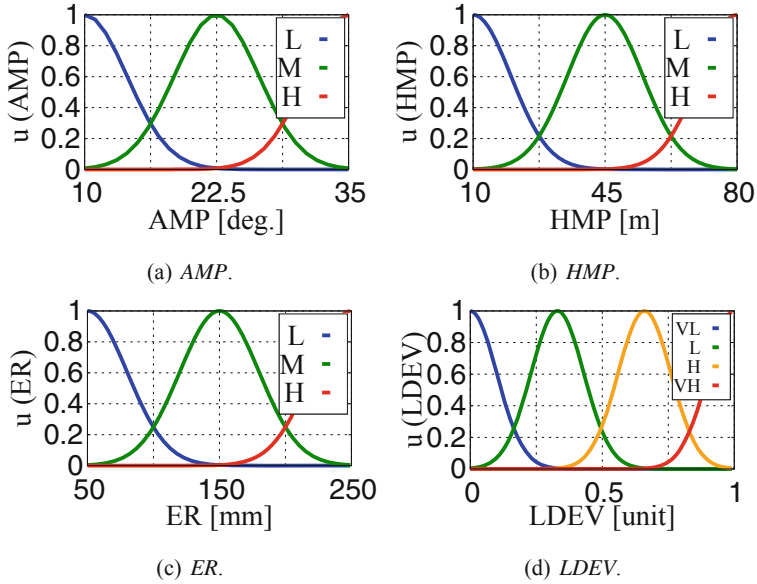


Fig. 2. Fuzzy membership functions.

Table 1. Fuzzy rule-base.

AMP	HMP	ER	LDEV	AMP	HMP	ER	LDEV	AMP	HMP	ER	LDEV
L	L	L	VL	M	L	L	VL	H	L	L	VL
L	L	M	VL	M	L	M	VL	H	L	M	VL
L	L	H	VL	M	L	H	VL	H	L	H	VL
L	M	L	VL	M	M	L	VL	H	M	L	VL
L	M	M	VL	M	M	M	L	H	M	M	H
L	M	H	L	M	M	H	H	H	M	H	VH
L	H	L	VL	M	H	L	VL	H	H	L	L
L	H	M	L	M	H	M	H	H	H	M	VH
L	H	H	L	M	H	H	VH	H	H	H	VH

2.1 LDR Estimation

The proposed Fuzzy-based system estimates the LDR considering data collection by WSN. It decides *Landslide Disasters Estimation Value (LDEV)* that indicates the risk of landslide disaster event. When *LDEV* value ($LDEV \in \mathbb{R}, 0.0 \leq LDEV \leq 1.0$) is close to 1.0, it indicates a higher value of LDR. The proposed system considers the following input parameters for deciding *LDEV*.

1. *Angle of Measurement Position of sensor [deg.] (AMP).*
2. *Height of Measurement Position of sensor [m] (HMP).*
3. *Effective Rainfall measured by rain gauge [mm] (ER).*

The ER parameter is decided by Eq. 1:

$$R_w = \sum_i^n 0.5^{i/T} R_i \quad (1)$$

where, R_w [mm] is ER , R_i [mm] is the rainfall i [hour] earlier and T [hour] is the half-life, which depends on the geology [25,26]. The n [hour] is the value for the ER analysis period. The ER is close to 0 as i increases. Therefore, the value of n is set until the influence of ER disappears. That is R_w is less than 10^{-3} .

The fuzzy rule-base and the membership functions are defined considering the landslide disasters warning areas by the Ministry of Land, Infrastructure, Transport and Tourism, in Japan. Table 1 shows the fuzzy rule-base. The Fig. 2(a), Fig. 2(b) and Fig. 2(c) show the input membership functions. Figure 2(d) shows the output membership function. As input and output terms we use: *Very-Low (VL)*, *Low (L)*, *Middle (M)*, *High (H)* and *Very-High (VH)*.

2.2 Danger Area Estimation Based on LDR

The danger area is estimated by LDR using elevation data from the Digital Elevation Model (DEM) of the Geospatial Information Authority of Japan (GSI). The proposed system decides the inclination angle of the target area from the elevation data. The proposed system decides LDR based on the elevation, inclination angle and effective rainfall on the days when landslide disasters have occurred in the target area in the past. The inclination angle S [27] is decided based on the surrounding of 8 grids cells centered at point E shown in Fig. 3 and Eqs. (2), (3) and (4). The S [deg.] is the angle, p is the weighted average from the west to east center point and q is the weighted average from the south to north center point.

$$S = \tan^{-1} \sqrt{p^2 + q^2} \quad (2)$$

$$p = \frac{(C + 2F + I) - (A + 2D + G)}{8\Delta x} \quad (3)$$

$$q = \frac{(G + 2H + I) - (A + 2D + C)}{8\Delta y} \quad (4)$$

Also, the elevation of each grid cell is from A to I , The Δx [m] and Δy [m] are the lengths of one side of the grid.

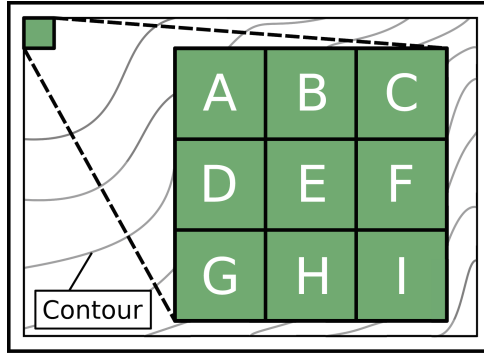


Fig. 3. Symbols for the grid cells.

3 Evaluation Results

3.1 LDR Estimation Results

For the estimation of LDR, we consider that the WSN obtains the input values for the proposed Fuzzy-based system. The HMP parameter is set to a constant value and AMP and ER are variable values. Figure 4 shows the visualization results of $LDEV$. Figure 4(a) and Fig. 4(b) show the results for HMP of 30.0 [m] and 60.0 [m], respectively. We can see that $LDEV$ increases with the increase of HMP .

For the evaluation, the proposed system decides $LDEV$ based on AMP , HMP and ER considering real landslide disasters sites. The AMP and HMP are decided from topographic data before the landslide disasters using the information from Geospatial Information Authority of Japan. In addition, ER is decided from the rainfall sensing data at the time of landslide disasters by the Japan Meteorological Agency. The half-life of ER is decided to be 72 [hour] considering the soil type of the target area. Table 2 shows the evaluation results of $LDEV$ for landslide disasters in Tsushima, Kita-ku, Okayama City, Okayama Prefecture, Japan and Furusato, Furusato-cho, Kagoshima City, Kagoshima Prefecture, Japan. In the case of Tsushima, the AMP is about 28 [deg.], HMP is about 72 [m] and ER is about 277 [mm]. The $LDEV$ value decided by Fuzzy-based system is 0.864 [unit]. In the case of Furusato, AMP is about 32 [deg.], HMP is about 94 [m] and ER is about 146 [mm]. The $LDEV$ value decided by Fuzzy-based system is 0.812 [unit]. The evaluation results show that the proposed system can estimate LDR according to the rainfall and topography parameters and by using real data collected on the wide areas by WSN.

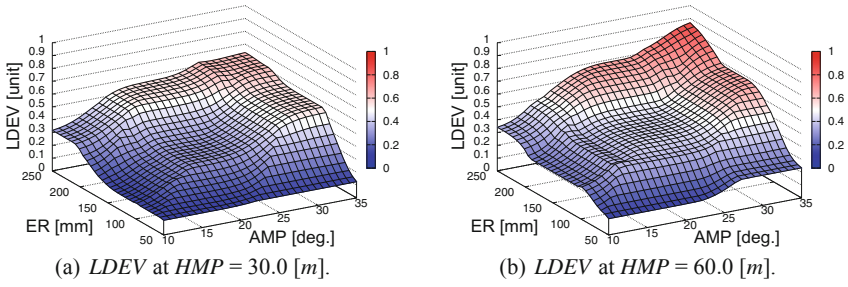
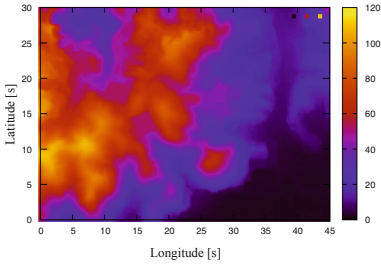


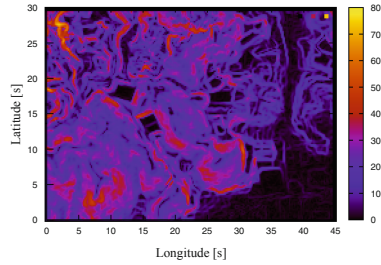
Fig. 4. Visualization results of $LDEV$.



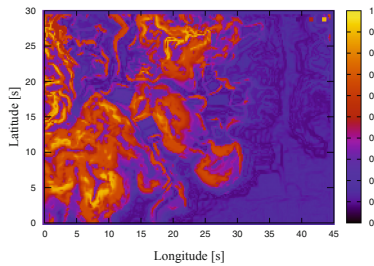
(a) Aerial view of the target mountain.



(b) Visualization results by height.



(c) Visualization results by angle.



(d) Visualization results by LDR

Fig. 5. Visualization results.

Table 2. Evaluation results of LDEV in the target area.

Lcation	<i>AMP</i> [deg.]	<i>HMP</i> [m]	<i>ER</i> [mm]	<i>LDEV</i> [unit]
Tsushima	28.4	72.4	277	0.864
Furusato	32.4	94.7	146	0.812

3.2 Evaluation Results for Danger Area Estimation Based on LDR

We show the visualization results in Fig. 5. Figure 5(a) shows the aerial view of the mountain used as target area. The visualization results of DEM elevation data are shown in Fig. 5(b), the decided inclination angles are shown in Fig. 5(c) and the occurrence danger areas based on LDR are shown in Fig. 5(d). The visualization results show the landslide disaster area can be visualized by using the *LDEV* estimated values from the height and inclination angle of the mountain.

4 Conclusions

In this paper, we proposed a Fuzzy-based system for the estimation of LDR considering the data collected by WSN and then to predict the landslide disaster areas based on LDR. Also, we evaluated the proposed system using the data from real landslide disaster sites. The evaluation results show that the proposed system can estimate LDR according to the rainfall and topography parameters and by using real data collected on the wide areas by WSN.

In the future, we would like to optimize the placement of sensors based on the predicted landslide disaster areas.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number JP20K19793.

References

1. Saito, H., et al.: Rainfall conditions, typhoon frequency and contemporary landslide erosion in Japan. *Geology* **42**(11), 999–1002 (2014)
2. Froude, M.J., et al.: Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazard.* **18**(8), 2161–2181 (2018)
3. Yoshida, M., et al.: Characteristics of disaster-related information in case of the heavy rain event of July 2018—a case study of Okayama, Hiroshima, and Ehime prefectures. *J. JSCE* **9**(1), 39–50 (2021)
4. Dou, J., et al.: An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. *Nat. Hazards* **78**(3), 1749–1776 (2015). <https://doi.org/10.1007/s11069-015-1799-2>
5. Ayalew, L., Yamagishi, H.: The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology* **65**(1–2), 15–31 (2005)
6. Ayalew, L., et al.: Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. *Landslides* **1**(1), 73–81 (2004)

7. Mucchi, L., et al.: A flexible wireless sensor network based on UltraWide band technology for ground instability monitoring. *Sensors* **18**(9), 2948 (2018)
8. Ramesh, M.V.: Real-time wireless sensor network for landslide detection. In: 2009 Third International Conference on Sensor Technologies and Applications, pp. 405–409. IEEE (2009)
9. Giri, P., et al.: Wireless sensor network system for landslide monitoring and warning. *IEEE Trans. Instrum. Meas.* **68**(4), 1210–1220 (2018)
10. Gu, X.B., et al.: The risk assessment of landslide hazards in Shiwangmiao based on intuitionistic fuzzy SetsTopsis model. *Nat. Hazards* **111**, 283–303 (2022)
11. Matsui, T., et al.: FPGA implementation of a fuzzy inference based quadrotor attitude control system. In: Proceedings of IEEE GCCE-2021, pp. 691–692 (2021)
12. Saito, N., et al.: Approach of fuzzy theory and hill climbing based recommender for schedule of life. In: Proceedings of LifeTech-2020, pp. 368–369 (2020)
13. Akhoondzadeh, M., Marchetti, D.: Developing a fuzzy inference system based on multi-sensor data to predict powerful earthquake parameters. *Remote Sens.* **14**(13), 3203 (2022)
14. Iwendi, C., et al.: Classification of COVID-19 individuals using adaptive neuro-fuzzy inference system. *Multimedia Syst.* **28**(4), 1223–1237 (2022)
15. Ozera, K., et al.: A fuzzy approach for secure clustering in MANETs: effects of distance parameter on system performance. In: Proceedings of IEEE WAINA-2017, pp. 251–258 (2017)
16. Yukawa, C., et al.: Evaluation of a fuzzy-based robotic vision system for recognizing micro-roughness on arbitrary surfaces: a comparison study for vibration reduction of robot arm. In: Proceedings of NBiS-2022, pp. 230–237 (2022)
17. Yukawa, C., et al.: Design of a fuzzy inference based robot vision for CNN training image acquisition. In: Proceedings of IEEE GCCE-2021, pp. 806–807 (2021)
18. Inaba, T., et al.: Performance evaluation of a QoS-aware fuzzy-based CAC for LAN access. *Int. J. Space Based Situated Comput.* **6**(4), 228–238 (2016)
19. Mukherjee, S., et al.: Evaluation of vertical accuracy of open source Digital Elevation Model (DEM). *Int. J. Appl. Earth Obs. Geoinf.* **21**, 205–217 (2013)
20. Claessens, L., et al.: DEM resolution effects on shallow landslide hazard and soil redistribution modelling. *Earth Surface Process. Land. J. Br. Geomorphol. Res. Group* **30**(4), 461–477 (2005)
21. Nagai, Y., et al.: A wireless sensor network testbed for monitoring a water reservoir tank: experimental results of delay. In: Proceedings of CISIS-2022, pp. 49–58 (2022)
22. Nagai, Y., et al.: A wireless sensor network testbed for monitoring a water reservoir tank: experimental results of delay and temperature prediction by LSTM. In: Proceedings of NBiS-2022, pp. 392–401 (2022)
23. Oda, T., et al.: Design and implementation of a simulation system based on deep Q-network for mobile actor node control in wireless sensor and actor networks. In: Proceedings of IEEE AINA-2017, pp. 195–200 (2017)
24. Yang, T., et al.: Impact of mobile sink nodes on performance of wireless sensor networks. *Int. J. Inf. Technol. Commun. Converg.* **2**(2), 155–170 (2012)
25. Hong, Y., et al.: The influence of intense rainfall on the activity of large-scale crystalline schist landslides in Shikoku Island, Japan. *Landslides* **2**(2), 97–105 (2005)
26. Vallet, A., et al.: Effective rainfall: a significant parameter to improve understanding of deep-seated rainfall triggering landslide—a simple computation temperature based method applied to Séchilienne unstable slope (French Alps). *Hydrol. Earth Syst. Sci. Discuss.* **10**(7), 8945–8991 (2013)
27. Horn, B.K.P.: Hill shading and the reflectance map. *Proc. IEEE* **69**(1), 14–47 (1981)



Human-Centered Protocols for Secure Data Management in Distributed Systems

Urszula Ogiela¹, Makoto Takizawa², and Lidia Ogiela¹(✉)

¹ AGH University of Science and Technology, 30 Mickiewicza Avenue, 30-059 Kraków, Poland
{ogiel, logiel}@agh.edu.pl

² Research Center for Computing and Multimedia Studies, Hosei University, 3-7-2, Kajino-cho,
Koganei-shi, Tokyo 184-8584, Japan
makoto.takizawa@computer.org

Abstract. In this paper a human-centered protocol will be presented for efficient and secure information management in distributed systems. Such techniques will be used in creation new security applications oriented on application of personal features in security solutions. Human-centered information management allow to facilitate data distribution in cloud infrastructure and distributed systems. Such techniques also allow to increase security of data management procedures.

Keywords: Human-centered protocols · Data security · Distributed systems · Data management

1 Introduction

Human-oriented data protection protocols are formed on the basis of analysis of human thought processes, the application of personal characteristics individual to each protocol user, and unique peculiarities characteristic of the person. Each user of a cryptographic data protection protocol can be assigned his or her individual characteristics, contained in personal sets of biometrics. These range from the well-known and widely used personal identification features, i.e. fingerprint, retinal scan, to the more complex ones, i.e. DNA code, fingerprint scan of both hands. However, regardless of the complexity of the solutions used, all biometric data can constitute an individual and unique kind of stamp marking each of its holders. This type of unique marking of the feature holder is extremely beneficial in the process of proper verification and identification [1–3]. This is because the uniqueness of the features is a guarantee that the system analyzing the data will not confuse or misidentify the holder of the features in question.

The use of individual biometrics in the process of personal identification and verification ensures that they are properly associated with their holder. However, solutions based on the use of more than one biometrics are increasingly being used. Such a solution is aimed at eliminating possible mistakes of the system performing the verification process.

Correct recognition at all stages of the identification process allows us to assume that the system has correctly linked the pattern with the holder of the biometric features. This

means correct verification of the system user. If one of the indicated biometrics is verified correctly and another is not, then additional identification of the user through additional analysis is required. The verification process may end in incorrect verification due to, for example, a bad or incomplete tag reading, reader error or system error. For example, the reading of the most popular biometric, which is the fingerprint, may encounter difficulties due to the fact that the fingerprint will be unreadable. In such cases, it is a good idea to use fingerprint analysis of all fingers because of the problems that can occur with reading a selected one.

Human-oriented protocols will be presented in this paper as those solutions that are designed to serve a particular type of data protection based on individual solutions [4, 5]. These models are open, which means that it is possible to modify them with new personal data/biometrics, the scope of which is inspired by the capabilities of computer systems, the area of application, the computing power of the machine units that conduct the analysis, and the speed of data processing.

2 Human-Centered Protocols for Secure Data Management

Human-oriented protocols are used for secure data management. These processes are implemented in a variety of information systems, with varying degrees of implementation, but nevertheless in the processes of information management it is necessary to secure information from unauthorized access, disclosure or declassification.

Data security processes, due to their diversity, are applied at various stages of data processing. Information that is not generally available, confidential, secret, protected, strategic, developmental, military, etc., is processed by persons authorized to possess it, but this does not exclude it from circulation in information systems, which are exposed to attempts to seize, hack or modify it.

Indeed, each type of the aforementioned information is of strategic importance and, due to this fact, is information desired by various entities and individuals. In order to protect this type of data, various solutions are used to secure it.

A new type of protocols that allow secure data management are human-oriented protocols. Their special feature is that algorithms based on individual personal characteristics are used in the process of information sharing, distribution, storage, and restoration.

The use of personal characteristics makes it possible to identify unambiguously the recipient of the information, and in the process of verification to properly identify its holder. A novelty aimed at enriching previously known solutions for secure data management are data sharing protocols based on human-centered solutions.

Human-oriented protocols, therefore, allow full verification of a participant in a secure data management protocol and verification of the part of all secret parts [6–8]. On the other hand, in the process of restoring the secret, they allow proper identification of all shadow holders. The human-centered protocol based on secure data management protocol is as follows:

Protocol 1. Human-centered protocol for secure data management

```
//definition of threshold scheme//
- determination of the number of participants ( $n$ )
  in the protocol
- definition of the number of shadows ( $m$ )
- definition of the type of data sharing -
  privileged/equal

//definition of the basic set of biometrics - random
selection of biometrics in the process of verification of
protocol participants//
- definition of a basic set of biometrics from which
  the system will randomly select a biometric on the
  basis of which each protocol participant will be
  verified
- definition of an additional set of biometrics for
  each protocol participant, to be used if
  verification based on the basic set of biometrics
  fails
- definition of the specific characteristics of
  protocol participants on the basis of which
  independent personal identification and
  verification will be possible

//definition of verification levels in the data management
protocol//
- definition of the stages of personal verification
- definition of processes for managing input
  information, classified data, shadows of shared
  secret
- definition of how to verify secret holders in the
  processes of secret sharing, distribution,
  processing, storage and restoration
```

3 Secure Data Management in Distributed Systems

Secure data management protocols based on the proposed solutions can be used in various areas of information management. One of them is distributed systems, operating not only within the selected structure but also outside it, and at different levels of management.

Distributed structures are characteristic of systems with a multilayer structure, whose data is sent outside the area of the entity where it was created/acquired. Transmitting them to external structures provides greater flexibility of the implemented processes and optimizes the cost of data storage and management. This type of solution allows to manage large collections of information from any level – both of the primary entity and of independent entities which are levels superior to the structure.

Data management at the primary level allows its distribution to the fog or cloud levels, from which the implementation of secret management processes is possible using the described solutions based on human-centered protocols.

These protocols applied at different levels of data management, while using human-centered solutions that guarantee proper processes for identifying and verifying protocol participants, make it possible to realize extensive and extensive processes of secure data management.

Indeed, the essence of this solution is the multi-level application of threshold schemes in the processes of data protection implemented with the participation of individual features of secret marking.

The multi-level nature of distributed structures is further enforced by the need for universal solutions that guarantee the effectiveness of the data protection process at different levels of the structure.

The ability to implement data security processes while securing and protecting data is due to the use of individual biometric tags of each part of the protected secret. An attempt to recreate this secret is as strongly protected as the entire shared information. This is also due to the use of biometrics to verify the parts of the secret being assembled.

The versatility of the proposed methods allows them to be widely used, from single structures where important data is processed and protected, to large entities operating in a changing environment and diverse external conditions.

4 Conclusions

The subject of this work was the possibility of using solutions based on human-centered protocols for secure data management in distributed structures. The possibility of developing such protocols stems from the versatility of threshold schemes, in which it is possible to incorporate shadow biometric marking processes arising from the division of protected/hidden information.

Biometric features as individual for each participant in the protocol allow to uniquely identify each of them, and their selection is random, making it more difficult to predict what solution will be chosen by the data protection system.

The implementation of such data protection processes can be carried out in classical data management structures, but can also be effectively implemented in distributed structures.

Acknowledgments. Research project supported by program “Excellence initiative – research university” for the AGH University of Science and Technology.

References

1. Menezes, A., van Oorschot, P., Vanstone, S.: Handbook of Applied Cryptography. CRC Press, Waterloo (2001)
2. Ogiela, M.R., Ogiela, U.: Secure information splitting using grammar schemes. In: Nguyen, N.T., Katarzyniak, R.P., Janiak, A. (eds.) New Challenges in Computational Collective Intelligence, vol. 244, pp. 327–336. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03958-4_28

3. Wojtowicz, W., Ogiela, M.R.: Digital images authentication scheme based on bimodal biometric watermarking in an independent domain. *J. Vis. Commun. Image Represent.* **38**, 1 (2016). <https://doi.org/10.1016/j.jvcir.2016.02.006>
4. Ogiela, M.R., Ogiela, L., Ogiela, U.: Biometric methods for advanced strategic data sharing protocols. In: 2015 9th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2015, pp. 179–183 (2015). <https://doi.org/10.1109/IMIS.2015.29>
5. Nakamura, S., Ogiela, L., Enokido, T., Takizawa, M.: Flexible synchronization protocol to prevent illegal information flow in peer-to-peer publish/subscribe systems. In: Barolli, L., Terzo, O. (eds.) *CISIS 2017. AISC*, vol. 611, pp. 82–93. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61566-0_8
6. Ogiela, L.: Transformative computing in advanced data analysis processes in the cloud. *Inf. Process. Manage.* **57**(5), 102260 (2020)
7. Ogiela, L., Ogiela, M.R.: Cognitive security paradigm for cloud computing applications. *Concurr. Comput. Pract. Exp.* **32**(8), e5316 (2020). <https://doi.org/10.1002/cpe.5316>
8. Yang, S.J.H., Ogata, H., Matsui, T., Chen, N.-S.: Human-centered artificial intelligence in education: seeing the invisible through the visible. *Comput. Educ. Artif. Intell.* **2**, 100008 (2021)



Multi-Version Concurrency Control to Reduce the Electric Energy Consumption of Servers

Tomoya Enokido¹(✉), Dilawaer Duolikun², and Makoto Takizawa³

¹ Faculty of Business Administration, Rissho University, 4-2-16, Osaki,
Shinagawa-ku, Tokyo 141-8602, Japan
`eno@ris.ac.jp`

² Department of Advanced Sciences, Faculty of Science and Engineering,
Hosei University, 3-7-2, Kajino-cho, Koganei-shi, Tokyo 184-8584, Japan

³ Research Center for Computing and Multimedia Studies, Hosei University,
3-7-2, Kajino-cho, Koganei-shi, Tokyo 184-8584, Japan
`makoto.takizawa@computer.org`

Abstract. The MVCC (Multi-Version Concurrency Control) is so far proposed to increase the concurrency of multiple conflicting transactions and the scalability of a distributed system. However, the larger number of transactions are concurrently performed, the larger amount of electric energy is consumed by servers in a system. In our previous studies, the EEMVTO (Energy-Efficient Multi-Version Timestamp Ordering) algorithm is proposed to not only reduce the total electric energy consumption of servers but also increase the throughput of a system by not performing meaningless write methods on each object. In this paper, the IEEMVTO (Improved EEMVTO) algorithm is newly proposed to furthermore reduce the total electric energy consumption of servers by not performing meaningless read methods in addition to meaningless write methods. The evaluation results show the total electric energy consumption of servers can be more reduced in the IEEMVTO algorithm than the EEMVTO algorithm.

Keywords: Multi-version concurrency control · Energy-Efficient Multi-Version Timestamp Ordering (EEMVTO) · Improved EEMVTO (IEEMVTO) algorithm · Object-based system · Transaction

1 Introduction

In current information systems, a huge number of IoT (Internet of Things) devices [1,2] are deployed in a system and each IoT device collects various types of data like temperature and humidity which are required by an application. A huge volume of data is gathered from these IoT devices in order to realize applications and the data gathered from IoT devices is encapsulated along with methods to manipulate data as an object [3] like database systems. An application is composed of multiple objects distributed to multiple physical servers in an

object-based system [3, 4, 6]. A transaction [7, 8] is an atomic sequence of methods to manipulate objects. In order to utilize an application service, a transaction is created on a client and issues methods supported by each target object. Multiple conflicting transactions have to be serialized [4, 6–11] to keep every object mutually consistent. The *MVCC* (*Multi-Version Concurrency Control*) [9, 10] is proposed to not only serialize conflicting transactions but also increase the concurrency of transactions and scalability of a system. In the MVCC, each read method is ensured to read the latest committed version of each object. In addition, each read method is not blocked by the other methods. As a result, the MVCC can increase the concurrency of transactions and the throughput of a system. In order to realize the MVCC, the *MVTO* (*Multi-Version Timestamp Ordering*) algorithm [9, 10] is proposed. However, the more number of transactions are issued in a system, the larger amount of electric energy is consumed by servers since every method issued to each target object is surely performed on each object. Hence, it is critical to discuss how to not only increase the concurrency of transactions and the throughput of a system but also reduce the total electric energy consumption of servers as discussed in Green computing systems [5, 6, 12–15].

In our previous studies, *meaningless write methods* [16] which are not required to be performed on each object are defined based on the precedent relation among transactions and the semantics of methods. Then, the *EEMVTO* (*Energy-Efficient Multi-Version Timestamp Ordering*) algorithm [16] is proposed to not only reduce the total electric energy consumption of servers but also increase the throughput of a system by not performing meaningless write methods on each object. In this paper, we newly introduce *meaningless read methods* which are not required to be performed on each object. Then, the *Improved EEMVTO* (*IEEMVTO*) algorithm is newly proposed to furthermore reduce the total electric energy consumption of servers and the execution time of each transaction by not performing both meaningless read and write methods. The IEEMVTO algorithm is evaluated in terms of the total electric energy consumption of servers and the average execution time of each transaction compared with the EEMVTO algorithm. Evaluation results show the total electric energy consumption of servers and the average execution time of each transaction in the IEEMVTO algorithm can be more reduced than the EEMVTO algorithm.

In Sect. 2, we present the system model and the MVTO algorithm. In Sect. 3, we propose the IEEMVTO algorithm. In Sect. 4, we evaluate the IEEMVTO algorithm compared with the EEMVTO algorithm.

2 System Model

2.1 Object-Based Systems

A system is composed of a cluster S of multiple servers s_1, \dots, s_n ($n \geq 1$) and clients interconnected in reliable networks. Let O be a set of objects o_1, \dots, o_m ($m \geq 1$) in the system. An object [3] is an unit of computation resource like a database. Each object o_h is an encapsulation of data d_h and methods to

manipulate data d_h in the object o_h . Each object o_h is allocated to a server s_t in the cluster S . Methods are classified into *read* (r) and *write* (w) methods in this paper. Write methods are furthermore classified into *full write* (fw) and *partial write* (pw) methods, i.e. $w \in \{fw, pw\}$. A full write method fully writes a whole data d_h in an object o_h . A partial write method writes only a part of data d_h in an object o_h .

2.2 Multi-Version Timestamp Ordering (MVTO) Algorithm

A *transaction* is an atomic sequence of methods [8]. A transaction T^i issues read (r) and write (w) methods to manipulate objects in the set O . Let \mathbf{T} be a set $\{T^1, \dots, T^k\}$ ($k \geq 1$) of transactions issued in a system. Multiple conflicting transactions are required to be *serializable* [7, 8] to keep all the objects mutually consistent. The *MVCC* (Multi-Version Concurrency Control) [9] is proposed to increase the concurrency of transactions and the throughput of a system. Let H be a schedule [9] of the transaction set \mathbf{T} . Each object o_h has a totally ordered set D_h of multiple versions d_h^1, \dots, d_h^l ($l \geq 1$) of data d_h . A totally ordered relation \ll_h ($\subseteq D_h^2$) shows an order of versions of data d_h of an object o_h written in a schedule H . $d_h^i \ll_h d_h^j$ means d_h^i is written before d_h^j in an object o_h . Let \ll be an union of version orders \ll_h for every data d_h in a schedule H , i.e. $\ll_h = \bigcup_{o_h \in O} \ll_h$. A transaction T^j *reads data from* another transaction T^i ($T^i \rightarrow_H T^j$) in a schedule H iff the transaction T^j reads a version d_h^i of an object o_h written by the transaction T^i . $T^i \parallel_H T^j$ iff neither $T^i \rightarrow_H T^j$ nor $T^j \rightarrow_H T^i$. A schedule H is $\langle \mathbf{T}, \rightarrow_H \rangle$ ($\subseteq \mathbf{T}^2$).

[One-Copy Serial]. A schedule $H = \langle \mathbf{T}, \rightarrow_H \rangle$ is *one-copy serial* [9] iff (if and only if) for every pair of different transactions T^i and T^j in \mathbf{T} , either $T^i \rightarrow_H T^j$, $T^j \rightarrow_H T^i$, or $T^i \parallel_H T^j$.

In an one-copy serial schedule $OH = \langle \mathbf{T}, \rightarrow_{OH} \rangle$ ($\subseteq \mathbf{T}^2$), if $T^i \rightarrow_H T^j$, $T^i \rightarrow_{OH} T^j$, and the relation, \rightarrow_{OH} is acyclic.

Let $r_t^i(d_h^j)$ be a read method issued by a transaction T^i to read a version d_h^j , which is written by a transaction T^j , of an object o_h on a server s_t . Let $w_t^i(d_h^i)$ be a write method issued by a transaction T^i to write a version d_h^i in an object o_h on a server s_t .

A *multi-version schedule MVS* is $\langle \mathbf{T}, \rightarrow_{MVS} \rangle$ ($\subseteq \mathbf{T}^2$) where for every pair of transactions T^i and T^j in \mathbf{T} , the following conditions hold:

- (1) If $T^i \rightarrow_{OH} T^j$, $T^i \rightarrow_{MVS} T^j$.
- (2) If T^i writes a version d_h^i , T^j reads a version d_h^k , and $T^i \rightarrow_{MVS} T^j$, $d_h^i \ll_h d_h^k$ or $d_h^i = d_h^k$.

[One-Copy Serializability]. A multi-version schedule $MVS = \langle \mathbf{T}, \rightarrow_{MVS} \rangle$ is *one-copy serializable* [9] iff for every pair of transactions T^i and T^j in \mathbf{T} , either $T^i \rightarrow_{MVS} T^j$, $T^j \rightarrow_{MVS} T^i$, or $T^i \parallel_{MVS} T^j$.

The *MVTO* (*Multi-Version Timestamp Ordering*) algorithm [9, 10] is proposed to make transactions one-copy serialize. Each transaction T^i is given an

unique timestamp $TS(T^i)$ which shows time when the transaction T^i is created. Suppose a transaction T^i issues a method op to manipulate an object o_h in a server s_t . In the MVTO algorithm, a method op issued by a transaction T^i is performed by the following procedure [9,10]:

1. If a method op is a read method $r_t^i(d_h^k)$, the read method op reads a version d_h^k written by a transaction T^k whose timestamp $TS(T^k)$ is the maximum in $TS(T^k) < TS(T^i)$.
2. If a method op is a write method $w_t^i(d_h^i)$, the write method op is rejected if a read method $r_t^j(d_h^k)$ is performed on the object o_h such that $TS(T^k) < TS(T^i) < TS(T^j)$. Otherwise, the write method $w_t^i(d_h^i)$ is performed.

By using the MVTO algorithm, each read method reads the latest committed version of an object o_h . In addition, each read method is not blocked by the other methods.

2.3 Data Access Model

Methods which are being performed and already terminate are *current* and *previous* at time τ , respectively. Let $RP_t(\tau)$ and $WP_t(\tau)$ be sets of current *read* (r) and *write* (w) methods on a server s_t at time τ , respectively. A notation $P_t(\tau)$ shows a set of current read and write methods on a server s_t at time τ , i.e. $P_t(\tau) = RP_t(\tau) \cup WP_t(\tau)$. Each read method $r_t^i(d_h^j)$ in a set $RP_t(\tau)$ reads a version d_h^j in an object o_h at rate $RR_t^i(\tau)$ [Byte/sec (B/sec)] at time τ . Each write method $w_t^i(d_h^i)$ in a set $WP_t(\tau)$ writes a version d_h^i in an object o_h at rate $WR_t^i(\tau)$ [B/sec] at time τ . Let $maxRR_t$ and $maxWR_t$ be the maximum read and write rates [B/sec] of read and write methods on a server s_t , respectively. The read rate $RR_t^i(\tau) (\leq maxRR_t)$ and write rate $WR_t^i(\tau) (\leq maxWR_t)$ are $dr_t(\tau) \cdot maxRR_t$ and $dw_t(\tau) \cdot maxWR_t$, respectively. Here, $dr_t(\tau)$ and $dw_t(\tau)$ are degradation ratios. $1 / (|RP_t(\tau)| + rw_t \cdot |WP_t(\tau)|)$ and $1 / (wr_t \cdot |RP_t(\tau)| + |WP_t(\tau)|)$, respectively, where $0 \leq rw_t \leq 1$ and $0 \leq wr_t \leq 1$. $0 \leq dr_t(\tau) \leq 1$ and $0 \leq dw_t(\tau) \leq 1$.

The *read laxity* $rl_t^i(\tau)$ [B] and *write laxity* $wl_t^i(\tau)$ [B] of methods $r_t^i(d_h^j)$ and $w_t^i(d_h^i)$ show the amount of data to be read and written in an object o_h by the methods $r_t^i(d_h^j)$ and $w_t^i(d_h^i)$ at time τ , respectively. Suppose that methods $r_t^i(d_h^j)$ and $w_t^i(d_h^i)$ start on a server s_t at time st_t^i . At time st_t^i , the read laxity $rl_t^i(\tau) = rb_h^j$ [B] where rb_h^j is the size of the version d_h^j in an object o_h . The write laxity $wl_t^i(\tau) = wb_h^i$ [B] where wb_h^i is the size of the version to be written in an object o_h . The read laxity $rl_t^i(\tau)$ and write laxity $wl_t^i(\tau)$ at time τ are $rb_h^j - \sum_{\tau=st_t^i}^{\tau} RR_t^i(\tau)$ and $wb_h^i - \sum_{\tau=st_t^i}^{\tau} WR_t^i(\tau)$, respectively.

2.4 Power Consumption Model of a Server

In our previous studies, the *PCS* model (*Power Consumption model for a Storage server*) [17] to perform storage and computation processes are proposed. Let

$E_t(\tau)$ be the electric power [W] of a server s_t at time τ . $maxE_t$ and $minE_t$ show the maximum and minimum electric power [W] of the server s_t , respectively. In this paper, we assume only read and write methods are performed on a server s_t . According to the PCS model [17], the electric power $E_t(\tau)$ [W] of a server s_t to perform multiple read and write methods at time τ is given as follows:

$$E_t(\tau) = \begin{cases} WE_t & \text{if } |WP_t(\tau)| \geq 1 \text{ and } |RP_t(\tau)| = 0. \\ WRE_t(\alpha) & \text{if } |WP_t(\tau)| \geq 1 \text{ and } |RP_t(\tau)| \geq 1. \\ RE_t & \text{if } |WP_t(\tau)| = 0 \text{ and } |RP_t(\tau)| \geq 1. \\ minE_t & \text{if } |WP_t(\tau)| = |RP_t(\tau)| = 0. \end{cases} \quad (1)$$

A server s_t consumes the minimum electric power $minE_t$ [W] if no method is performed on the server s_t , i.e. the electric power in the idle state of the server s_t . The server s_t consumes the electric power RE_t [W] if at least one r method is performed on the server s_t . The server s_t consumes the electric power WE_t [W] if at least one w method is performed on the server s_t . The server s_t consumes the electric power $WRE_t(\alpha)$ [W] $= \alpha \cdot RE_t + (1 - \alpha) \cdot WE_t$ [W] where $\alpha = |RP_t(\tau)| / (|RP_t(\tau)| + |WP_t(\tau)|)$ if both at least one r method and at least one w method are concurrently performed. Here, $minE_t \leq RE_t \leq WRE_t(\alpha) \leq WE_t \leq maxE_t$. The total electric energy $TEE_t(\tau_1, \tau_2)$ [J] of a server s_t from time τ_1 to τ_2 is $\sum_{\tau=\tau_1}^{\tau_2} E_t(\tau)$. The processing electric power $PEP_t(\tau)$ [W] of a server s_t at time τ is $E_t(\tau) - minE_t$. The total processing electric energy $TPEE_t(\tau_1, \tau_2)$ of a server s_t from time τ_1 to τ_2 is given as $TPEE_t(\tau_1, \tau_2) = \sum_{\tau=\tau_1}^{\tau_2} PEP_t(\tau)$.

3 Improved EEMVTO (IEEMVTO) Algorithm

3.1 Meaningless Methods

Let MH_h be a *local schedule* of methods which are performed on an object o_h in a multi-version schedule MH . A method op^1 of a transaction T^1 *locally precedes* another method op^2 of a transaction T^2 in a local schedule MH_h ($op^1 \rightarrow_{MH_h} op^2$) iff $T^1 \rightarrow_{MH} T^2$ and op^1 is performed before op^2 on an object o_h . Suppose a partial write method $pw^i(d_h^i)$ issued by a transaction T^i locally precedes another full write method $fw^j(d_h^j)$ issued by a transaction T^j in a local schedule MH_h ($pw^i(d_h^i) \rightarrow_{MH_h} fw^j(d_h^j)$) on an object o_h . Here, the partial write method $pw^i(d_h^i)$ is not required to be performed on the object o_h if the full write method $fw^j(d_h^j)$ is surely performed on the object o_h just after the partial write method $pw^i(d_h^i)$, i.e. the full write method $fw^j(d_h^j)$ can *absorb* the partial write method $pw^i(d_h^i)$.

[Absorption of Write Methods]. A full write method op^1 *absorbs* another partial or full write method op^2 in a local subschedule MH_h on an object o_h iff one of the following conditions is hold:

1. $op^2 \rightarrow_{MH_h} op^1$ and there is no read method op' such that $op^2 \rightarrow_{MH_h} op' \rightarrow_{MH_h} op^1$.
2. op^1 absorbs op^3 and op^3 absorbs op^2 for some method op^3 .

[Absorption of Read Methods]. A read method op^1 absorbs another read method op^2 in a local subschedule H_h of an object o_h iff one of the following conditions is hold:

1. $op^1 \rightarrow_{H_h} op^2$ and there is no write method op' such that $op^1 \rightarrow_{H_h} op' \rightarrow_{H_h} op^2$.
2. op^1 absorbs op^3 and op^3 absorbs op^2 for some method op^3 .

[Meaningless Methods]. A method op is *meaningless* iff the method op is absorbed by another method op' in the local subschedule MH_h on an object o_h .

3.2 IEEMVTO Algorithm

In this paper, the *IEEMVTO* (*Improved EEMVTO*) algorithm is newly proposed to furthermore reduce not only the total electric energy consumption of a cluster of servers but also the average execution time of each transaction by not performing meaningless read and write methods on each object. In this paper, we assume transactions are serialized based on the MVTO algorithm [9,10].

Suppose a read method $r_t^i(d_h^k)$ issued by a transaction T^i is performed on the object o_h as shown in Fig. 1. A transaction T^j issues a read method $r_t^j(o_h^k)$ to the object o_h while the read method $r_t^i(d_h^k)$ is being performed on the object o_h . In the MVTO algorithm, the read method $r_t^j(o_h^k)$ is performed on the object o_h as soon as the object o_h receives the read method $r_t^j(o_h^k)$. In the IEEMVTO algorithm, the read method $r_t^j(o_h^k)$ is meaningless since the read method $r_t^i(o_h^k)$ issued by the transaction T^i is being performed on the object o_h and the read method $r_t^i(o_h^k)$ absorbs the read method $r_t^j(o_h^k)$. Hence, the read method $r_t^j(o_h^k)$ is not performed on the object o_h and a result obtained by performing the read method $r_t^i(o_h^k)$ is sent to a pair of transactions T^i and T^j .

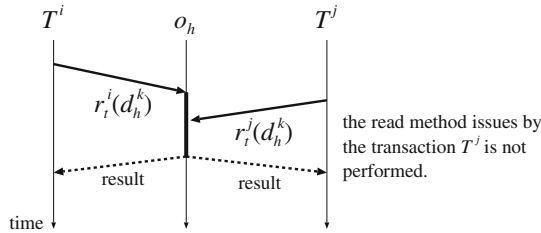


Fig. 1. A meaningless read method.

Suppose a transaction T^i issues a partial write method $pw_t^i(d_h^i)$ to an object o_h allocated to a server s_t as shown in Fig. 2. In the MVTO algorithm, the partial write method $pw_t^i(d_h^i)$ is performed on the object o_h as soon as the object o_h receives the partial write method $pw_t^i(d_h^i)$. In the EEMVTO algorithm, the

object o_h sends a termination notification of the partial write method $pw_t^i(d_h^i)$ to the transaction T^i as soon as the object o_h receives the partial write method $pw_t^i(d_h^i)$. However, the partial write method $pw_t^i(d_h^i)$ is not performed until the object o_h receives a method op which is performed just after the partial write method $pw_t^i(d_h^i)$ on the object o_h , i.e. the partial write method $pw_t^i(d_h^i)$ is delayed. Suppose a transaction T^j issues a full write methods $fw_t^j(d_h^j)$ to the object o_h after the transaction T^i commits. Here, the partial write method $pw_t^i(d_h^i)$ issued by the transaction T^i is meaningless since the full write method $fw_t^j(d_h^j)$ issued by the transaction T^j absorbs the partial write method $pw_t^i(d_h^i)$ on the object o_h . Hence, the full write method $fw_t^j(d_h^j)$ can be performed on the object o_h without performing the partial write method $pw_t^i(d_h^i)$. This means that the meaningless write method $pw_t^i(d_h^i)$ is not performed on the object o_h .

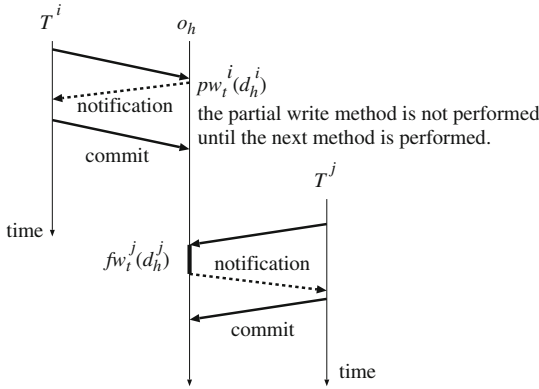


Fig. 2. Omission of a meaningless write method.

Suppose a transaction T^j issues a read method $r_t^j(d_h^i)$ after another transaction T^i commits. Here, the partial write method $pw_t^i(d_h^i)$ issued by the transaction T^i has to be performed before the read method $r_t^j(d_h^i)$ is performed since the read method $r_t^j(d_h^i)$ has to read a version d_h^i written by the partial write method $pw_t^i(d_h^i)$.

Let $o_h.Cr$ be a read method $r_t^i(d_h^k)$ issued by a transaction T^i , which is being performed on a object o_h . A notation $o_h.Dw$ is a write method $w_t^i(d_h^i)$ issued by a transaction T^i to write data d_h^i of an object o_h in a server s_t , which is waiting for a method op to be performed on the object o_h after $w_t^i(d_h^i)$. Suppose a transaction T^i issues a method op to an object o_h . In the IEEMVTO algorithm, the method op is performed on the object o_h by the following IEEMVTO procedure:

```

IEEMVTO( $op$ ) {
  if  $op = r$ , { /*  $op$  is a read method. */
    if  $o_h.Dw = \phi$ , {
      if  $o_h.Cr = \phi$ , {

```

```

     $o_h.Cr = op(d_h^k);$ 
    perform( $op(d_h^k)$ ); /*  $d_h^k$  is the latest committed data. */
     $o_h.Cr = \phi;$ 
  }
  else a result of  $o_h.Cr$  is sent to a transaction  $T^i$ ;
}
else {
  perform( $o_h.Dw$ );
   $o_h.Dw = \phi;$ 
   $o_h.Cr = op(d_h^k);$ 
  perform( $op(d_h^k)$ ); /*  $d_h^k$  is the latest committed data. */
   $o_h.Cr = \phi;$ 
}
}
else { /*  $op$  is a write method. */
  if  $o_h.Dw = \phi, o_h.Dw = op(d_h^i);$ 
  else { /*  $o_h.Dw \neq \phi$  */
    if  $op(d_h^i)$  absorbs  $o_h.Dw, o_h.Dw = op(d_h^i);$  /*  $o_h.Dw$  is not performed. */
    else {
      perform( $o_h.Dw$ );
       $o_h.Dw = op(d_h^i);$ 
    }
  }
}
}
}

```

In the IEEMVTO algorithm, the total electric energy consumption of a cluster S of servers can be furthermore reduced than the EEMVTO algorithm since the number of read and write methods performed on each object can be more reduced. In addition, the computation resources which are used to perform meaningless read and write methods can be used to perform the other methods in each server s_t . As a result, the execution time of each transaction can be more reduced in the IEEMVTO algorithm than the EEMVTO algorithm. This means that the throughput of a system can increase in the IEEMVTO algorithm than the EEMVTO algorithm.

4 Evaluation

4.1 Environment

We evaluate the IEEMVTO algorithm in terms of the total processing electric energy of a cluster S of homogeneous servers and the average execution time of each transaction compared with the EEMVTO algorithm [16]. The cluster S of servers is composed of ten homogeneous servers s_1, \dots, s_{10} ($n = 10$), where every server s_t ($t = 1, \dots, 10$) follows the same data access model and power consumption model. Parameters of each server s_t are shown in Table 1, which

are obtained based on the experimentations [17]. There are thirty objects o_1, \dots, o_{30} in a system. The size of data in each object o_h is randomly selected between 50 and 100 [MByte]. Each object o_h supports *read* (r), *full write* (fw), and *partial write* (pw) methods. Each object is randomly allocated to a server s_t in the cluster S .

Table 1. Homogeneous cluster S of servers ($t = 1, \dots, 10$)

Server s_t	$maxRR_t$	$maxWR_t$	rw_t	wr_t	$minE_t$	WE_t	RE_t
s_t	80 [MB/sec]	45 [MB/sec]	0.5	0.5	39 [W]	53 [W]	43 [W]

The number nt ($0 \leq nt \leq 500$) of transactions are issued to manipulate objects. Each transaction issues three methods randomly selected from one-hundred fifty methods on the fifty objects. The total amount of data of an object o_h is fully written by each full write (fw) method. On the other hand, a half size of data of an object o_h is written and read by each partial write (pw) and read (r) methods, respectively. The starting time of each transaction T^i is randomly selected in a unit of one second between 1 and 360 [sec].

4.2 Total Processing Electric Energy Consumption

Figure 3 shows the total processing electric energy consumption [KJ] of the cluster S of servers to perform the number nt of transactions in the IEEMVTO and EEMVTO algorithms. For $0 \leq nt \leq 500$, the total processing electric energy consumption of the cluster S of servers can be more reduced in the IEEMVTO algorithm than the EEMVTO algorithm. In the IEEMVTO algorithm, meaningless read and write methods are not performed on each object. As a result, the total processing electric energy consumption of the cluster S of servers can be more reduced in the IEEMVTO algorithm than the EEMVTO algorithm.

4.3 Average Execution Time of Each Transaction

Figure 4 shows the average execution time [sec] of the nt transactions in the IEEMVTO and EEMVTO algorithms. In the IEEMVTO and EEMVTO algorithms, the average execution time increases as the total number nt of transactions increases since more number of transactions are concurrently performed. For $0 < nt \leq 500$, the average execution time of each transaction can be more reduced in the IEEMVTO algorithm than the EEMVTO algorithm. In the IEEMVTO algorithm, each transaction can commit without waiting for performing meaningless methods. Hence, the average execution time of each transaction is shorter in the IEEMVTO algorithm than the EEMVTO algorithm.

Following the evaluation, the total processing electric energy consumption of a homogeneous cluster S of servers and the average execution time of each

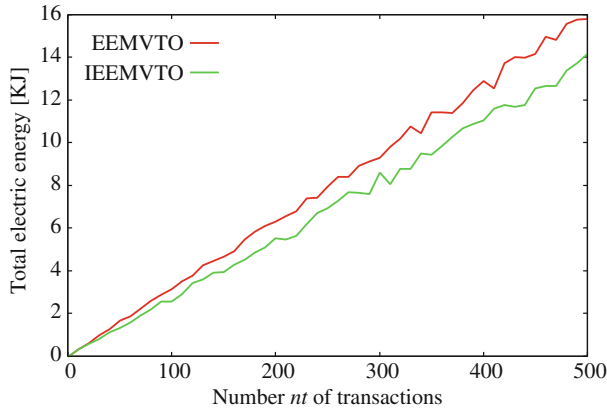


Fig. 3. Total processing electric energy consumption [KJ] of a cluster S of servers.

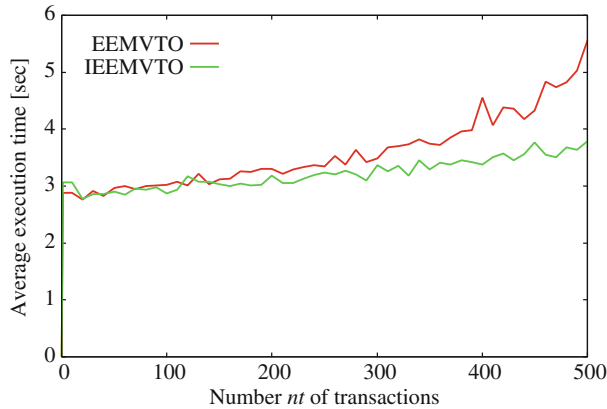


Fig. 4. Average execution time [sec] of each transaction.

transaction can be more reduced in the IEEMVTO algorithm than the EEMVTO algorithm. Hence, the IEEMVTO algorithm is more useful than the EEMVTO algorithm.

5 Concluding Remarks

In this paper, we newly proposed the IEEMVTO algorithm to reduce not only the total processing electric energy consumption of a cluster of servers but also the average execution time of each transaction by not performing meaningless read and write methods. We evaluated the IEEMVTO algorithm compared with the EEMVTO algorithm. The evaluation results showed the total processing electric energy consumption of a cluster of servers and the average execution time of each transaction can be more reduced in the IEEMVTO algorithm than

the EEMVTO algorithm. Following the evaluation, the IEEMVTO algorithm is more useful than the EEMVTO algorithm.

References

1. Nakamura, S., Enokido, T., Takizawa, M.: Implementation and evaluation of the information flow control for the Internet of Things. *Concurr. Comput. Practice Exp.* **33**(19), e6311 (2021)
2. Enokido, T., Takizawa, M.: The redundant energy consumption laxity based algorithm to perform computation processes for IoT services. *Internet Things* **9** (2020). <https://doi.org/10.1016/j.iot.2020.100165>
3. Object Management Group Inc.: Common object request broker architecture (CORBA) specification, version 3.3, Part 1 - interfaces (2012). <https://www.omg.org/spec/CORBA/3.3/Interfaces/PDF>
4. Tanaka, K., Hasegawa, K., Takizawa, M.: Quorum-based replication in object-based systems. *J. Inf. Sci. Eng.* **16**(3), 317–331 (2000)
5. Enokido, T., Duolikun, D., Takizawa, M.: An energy-efficient quorum-based locking protocol by omitting meaningless methods on object replicas. *J. High Speed Netw.* **28**(3), 181–203 (2022)
6. Enokido, T., Duolikun, D., Takizawa, M.: Energy-efficient concurrency control by omitting meaningless write methods in object-based systems. In: *Proceedings of the 36th International Conference on Advanced Information Networking and Applications (AINA-2022)*, pp. 129–139 (2022)
7. Gray, J.N.: Notes on data base operating systems. In: Bayer, R., Graham, R.M., Seegmüller, G. (eds.) *Operating Systems*. LNCS, vol. 60, pp. 393–481. Springer, Heidelberg (1978). https://doi.org/10.1007/3-540-08755-9_9
8. Bernstein, P.A., Hadzilacos, V., Goodman, N.: *Concurrency Control and Recovery in Database Systems*. Addison-Wesley, Boston (1987)
9. Bernstein, P.A., Goodman, N.: Multiversion concurrency control - theory and algorithms. *ACM Trans. Database Syst.* **8**(4), 465–483 (1983)
10. Reed, D.: Naming and synchronization in a decentralized computer system. Technical report, MIT/LCS/TR-205, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (1978). <https://hdl.handle.net/1721.1/16279>
11. Garcia-Molina, H., Barbara, D.: How to assign votes in a distributed system. *J. ACM* **32**(4), 814–860 (1985)
12. Natural Resources Defense Council (NRDC): Data center efficiency assessment - scaling up energy efficiency across the data center industry. Evaluating key drivers and barriers (2014). <https://www.nrdc.org/energy/files/data-center-efficiency-assessment-IP.pdf>
13. Enokido, T., Duolikun, D., Takizawa, M.: Energy consumption laxity-based quorum selection for distributed object-based systems. *Evol. Intel.* **13**(1), 71–82 (2018). <https://doi.org/10.1007/s12065-018-0157-1>
14. Enokido, T., Duolikun, D., Takizawa, M.: The improved redundant active time-based (IRATB) algorithm for process replication. In: Barolli, L., Woungang, I., Enokido, T. (eds.) *AINA 2021. LNNS*, vol. 225, pp. 172–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-75100-5_16

15. Enokido, T., Duolikun, D., Takizawa, M.: The redundant active time-based algorithm with forcing meaningless replica to terminate. In: Barolli, L., Yim, K., Enokido, T. (eds.) CISIS 2021. LNNS, vol. 278, pp. 206–213. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79725-6_20
16. Enokido, T., Duolikun, D., Takizawa, M.: Energy-efficient multi-version concurrency control (EEMVCC) for object-based systems. accepted for publication. In: Barolli, L., Miwa, H., Enokido, T. (eds.) NBiS 2022. LNNS, vol. 526. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-14314-4_2
17. Sawada, A., Kataoka, H., Duolikun, D., Enokido, T., Takizawa, M.: Energy-aware clusters of servers for storage and computation applications. In: Proceedings of the 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016), pp. 400–407 (2016)



A Study on Increasing Simultaneous Transmissions After Extended RTS/CTS Handshake on Full-duplex Wireless LANs

Hikari Hashimoto and Tetsuya Shigeyasu^(✉)

Graduate School of Comprehensive Scientific Research, Prefectural University of Hiroshima,
Hiroshima, Japan

r222004zc@ed.pu-hiroshima.ac.jp, sigeyasu@pu-hiroshima.ac.jp

Abstract. Recently, with the rapid development of the Internet of Things (IoT) technology, wide variety of objects in our surrounding environment newly connected to a network. In order to provide continuous network services, the limited communication resources must be utilized effectively.

In order to increase network capacity, full-duplex wireless communication allowing a node to simultaneously transmit while receiving on the same communication channel has become more attractive. Compared to the half-duplex wireless communication where a node can either transmit or receive at once, full-duplex wireless communication can achieve as up to twice as the communication capacity. However, in order to maximize the use of full-duplex wireless communication, appropriate scheduling on simultaneous transmissions and selecting the appropriate destinations avoiding collisions with the other transmissions, are required.

In this paper, we propose a method to improve communication throughput by selecting multiple secondary senders from primary sender's neighbors that never interfere with the other primary transmissions. In addition, our proposal make new transmission immediately after the extended RTS/CTS handshake for protecting the new full-duplex transmissions.

1 Introduction

Recently, with the rapid development of the IoT [1] technology, we can monitor and drive a lot of devices in the real spaces with sensors and actuators. In order to connect wide variety of objects in our surrounding environment to the network newly and provide continuous network services, the limited communication resources must be utilized effectively.

It is obviously that total network bandwidth can be increased by additional communication channel which is independent from the other exiting channels. In the case of wired network, it is easy to add such channel by adding new cable to the network. This is because, basically, different cable does not interfere the other cable separated physically, each other. On the other hand, in the case of wireless communications, it is hard to increase network capacity because of a difficulty of remove interferences among neighboring communication nodes in the same radio frequency. However, due

to the easy placement without any cabling, wireless connections are strongly requested to utilize the ability of IoT devices. In wireless communications, collisions are one of the reasons of performance reduction. Especially, on Carrier Sense Multiple Access (CSMA), in which data packets are sent immediately after carrier sensing, collision occurs when a node around the destination node starts at last one new transmission. Therefore, it is effective to use RTS/CTS (Request To Send/Clear To Send) exchange prior to DATA transmission in order to suppress interference transmissions from neighbor node of receiver. Similarly, in the case of full-duplex wireless communication, collision avoidance is also essential to improve performance.

One of the strong candidate methods for increasing a channel capacity in one band, is full-duplex wireless communication [2] allowing a node to simultaneously transmit while its receiving on the same communication channel. Compared to the half-duplex wireless communication where a node can either transmit or receive at once, full-duplex wireless communication can achieve as up to twice as the communication capacity. However, in order to maximize the use of full-duplex wireless communication, appropriate scheduling on simultaneous transmission [3–5] and selecting the appropriate destinations avoiding collisions among the other transmissions [6, 7], are required. At the same time, it is necessary to develop control packets and communication protocols for full-duplex wireless communication.

In this paper, we propose a method to improve communication throughput by selecting multiple secondary senders from primary sender's neighbors that never interfere with other primary transmissions. In addition, our proposal make new transmissions immediately after the extended RTS/CTS handshake for protecting the new full-duplex transmissions. The result of computer simulations confirms that our proposal effective increases the network capacity.

2 Full-duplex Wireless Transmission

Full-duplex wireless communication can perform transmission and reception simultaneously on the same channel. In this section, we describe the characteristics of full-duplex wireless communication. Hereafter, the primary transmission refers to the first communication in full-duplex wireless communication, and the secondary communication refers to the communication started synchronously with the primary transmission.

2.1 Full-duplex Transmission Fashions

Full-duplex communications can be divided into bi-directional full-duplex communications and relay full-duplex communications according to the differences of the transmission and reception node type [8]. In addition, the latter relay full-duplex communication can be further divided into two types depending on the difference of the secondary destination node. In the following description, a primary transmitter and a primary receiver are indicated as PT and PR, and a secondary transmitter and a secondary receiver are indicated as ST and SR, respectively.

Bi-directional Full-duplex Transmissions. Bi-directional full-duplex transmissions can be performed if and only if the any two neighboring nodes have at least one packet destined to each other. On bi-directional full-duplex communication, PR receives a primary packet and sends a secondary packet to PT. Therefore, as shown in Fig. 1, PT and SR, PR and ST are the same node.

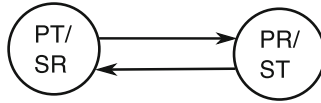


Fig. 1. Bi-directional full-duplex transmission

Relay Full-duplex Communication. Relay full-duplex communication is performed by three nodes. This communication can be further classified into PR-based relay full-duplex communication and PT-based relay full-duplex communication based on the difference of the role of ST.

PR-based relay full-duplex communication is performed when PR has packets other than PT. Therefore, as shown in Fig. 2, PR and ST are the same node, but PT and SR are different nodes.

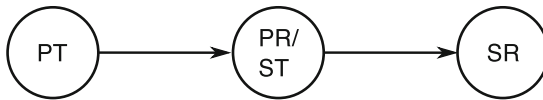


Fig. 2. PR-based full-duplex transmission.

On the other hand, the latter PT-based relay full-duplex communication is performed when a neighbor other than PR hearing the primary transmission packet destined to PT. Therefore, as shown in Fig. 3, PT and SR are the same node, but PR and ST are different nodes.

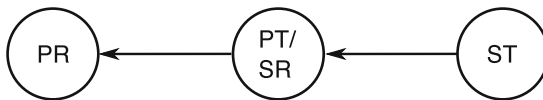


Fig. 3. PT-based full-duplex transmission.

Relay full-duplex communication has an advantage that it can be applied to nodes employing only half-duplex communication. In Fig. 2 and 3, only the center node can both transmits and receives simultaneously, while the nodes on the left and opportunity only transmit or receive.

2.2 Challenges of Full-duplex Wireless Transmission

One of the challenges of full-duplex wireless communication is the interference [9–12]. As shown in Fig. 4, self-interference and inter-user interference will be generated during full-duplex wireless communications.

Self-interference is induced by its transmission wave on a receiving node. Due to the property of its transmission wave is known, self-interference can be reduced by advanced signal processing techniques using analog and digital circuits [9].

On the other hand, inter-user interference is caused by transmission waves from the other nodes. In the case of bi-directional full-duplex communication, new transmissions within the transmission range of both PT and ST are suppressed, so inter-user interference rarely occur. In the case of relay full-duplex communication, if the SR exists within the communication range of the PT, as shown in Fig. 4, the primary transmission interfered and the desired secondary transmission will be degraded. It is difficult to remove the inter-user interference from unknown multiple nodes than self-interference.

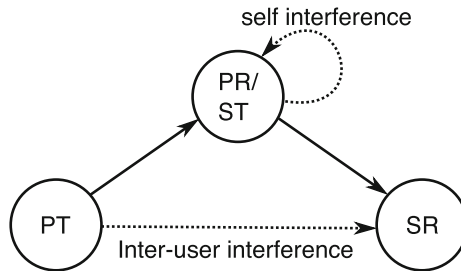


Fig. 4. Self-interference and inter-user interference in PR-based relay full-duplex communications.

3 Method for Selecting Multiple Nodes for Synchronous Transmission

In this section, we propose a method selecting multiple nodes for synchronous transmissions together with prior started PR-based relay full-duplex transmission. Here, nodes for synchronous transmissions are selected from PT's neighbor nodes. These nodes will be selected according to relationship of node connections each other. Selected nodes are ensured that receivers of those selected transmitters never receive interference from the other synchronous transmitters. In addition, nodes selected as synchronous transmitters are not required to implement a full-duplex function. Hence, our proposal can select half-duplex nodes as synchronous transmitters. The proposed method aims to increase the transmission opportunities by increasing the number of synchronous transmitters and improving throughput by avoiding collisions between each transmission nodes.

3.1 Transmission Procedure

Figures 5, 6 show the examples of topology and transmission sequence of our proposal. In parallel with PDATA, SynT sends SynR DATAsyn. In Fig. 5, the solid line indicates the suppression range of PRTS, the double line indicates the suppression range of PCTS, and the dotted line indicates the suppression range of SCTS. PRTS includes the node ID of PT, PR and candidate transmission nodes list, PCTS includes the node ID of PT, PR, ST and SR, and SCTS includes the node ID of ST and SR.

Each node has random backoff, independently when a transmission request occurs, and PT that has the transmission opportunity sends PRTS with the node ID at the head of the transmission queue as PR. Nodes receiving PRTS proceed on the basis of its node type as follows:

1. PR sends back PCTS when it is idle
2. candidate transmission nodes set NAV until PDATA start time
3. another nodes set NAV until PDATA transmission completion time

Nodes receiving PCTS proceeds on the basis of its node type as follows:

1. PT can send PDATA after SCTS transmission completion time
2. SR sends back SCTS
3. another nodes set NAV until SCTS transmission completion time

Nodes receiving SCTS proceeds on the basis of its node type as follows:

1. ST can send SDATA synchronously with PDATA
2. another nodes set NAV until the completion time of SDATA transmission

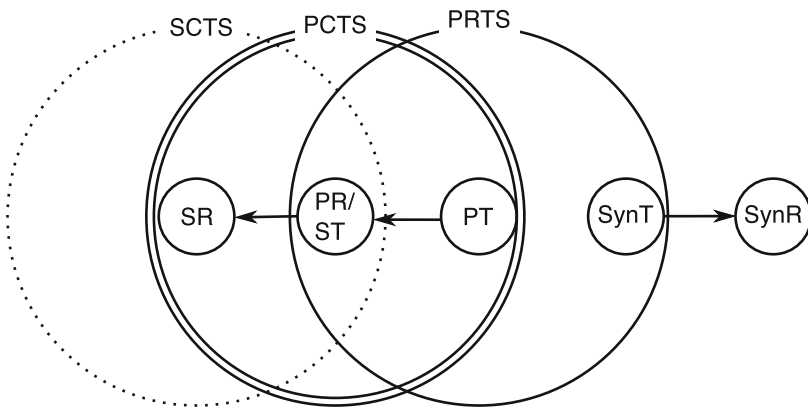


Fig. 5. Example topology.

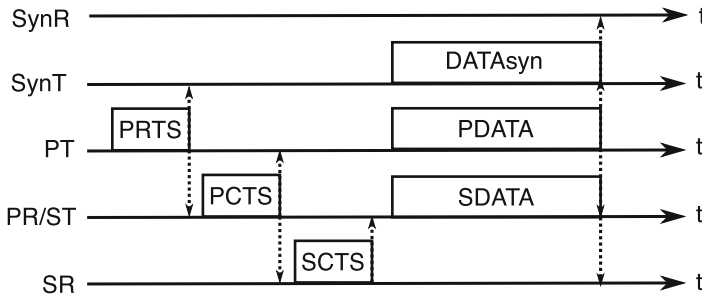


Fig. 6. Transmission sequence.

3.2 Selection Algorithm of Candidate Transmitters

In our proposal, synchronous transmission nodes will be selected from PT's neighbor nodes for avoiding collision between other transmissions. The procedures of making such list are described below:

1. Make the set S be the collection of neighbor PT
2. Pick up nodes fulfilling the following conditions:
 - a. not connected to PR
 - b. not shared the destination with PR
3. Calculate *influence* of nodes selected on step 1, and add a node to candidate node list if it has a minimum value of *influence*.
4. Remove node from S , fulfilling the following conditions:
 - a. candidate node selected on step3
 - b. neighbor of candidate node
 - c. node sharing destination with PR

In the step 2 on the above procedure, value of *influence* is a number of nodes that are 2 hops away from PT and neighbors of the candidate node. We propose that increase the number of synchronous transmissions by selecting the node with the smallest *influence* to the neighbors.

3.3 Method for Selecting Destination of Synchronous Transmission

PT sends packet as a primary transmission to the node recorded in a packet on a head of its queue. ST also behaves as PR, sends packet as secondary transmission to the node outside the PT's transmission range because of avoiding the collision on the PR. SynT described by the candidate list also determines the SynR outside the PT's transmission range. If neither the ST nor the SynT has a transmission queue that meets the conditions, it postpones own new transmission until the PDATA transmission end time. When it have no transmission queue meeting the required conditions even if it is designated as the ST or the SynT.

4 Performance Evaluation

This section reports the results of performance evaluation of the proposal. Simulation parameters are shown in Table 1. PT is placed in the center, and the other nodes are placed randomly in a simulation field. Full-duplex wireless communication will be performed only when the center node gets the transmission opportunity. When other node gets a floor, it performs half-duplex wireless communication after RTS/CTS exchange.

Table 1. Simulation parameters

Parameter	Value
Transmission speed	1 Mbps
Communication range	100 m
Simulation period	10.0 s
Simulation field	500 m \times 500 m
Packet arrival process	Poisson distribution

The following six methods comparison methods in the performance evaluation.

- SF w/o RTS
- SFS w/o RTS
- SFSC w/o RTS
- SF w/ RTS
- SFS w/ RTS
- SFSC w/ RTS

w/o RTS does not use RTS/CTS exchange. In those methods, center node that has acquired the transmission opportunity by carrier sense, starts data transmission immediately. In methods named with “w/ RTS”, when the center node has acquired the transmission opportunity, it performs PRTS/PCTS/SCTS exchange before each data packet transmission. SF (Single relay Full-duplex transmission) is a PR-based relay full-duplex communication. SFS (SF with Synchronous transmission) performs synchronous transmission by proposed algorithm without step 2b and step 4c. SFS increases total number of transmissions than SF. SFSC (SFS with Canceling collision) performs synchronous transmission complete proposed algorithm. Applying to the algorithm step 2b and step 4c, the nodes that share SR (SynR) with selected ST (SynT) are suppressed new transmissions. Hence, by evaluating SFSC, we can confirm the effects of avoiding overlapping selection of SR (SynR).

4.1 Characteristics of Throughput Performance

Figure 7 shows the characteristics of throughput. In this figure, throughput performance indicates the total successful transmissions relating to the primary transmission by center placed node. Even though the transmission requests are arrived at all nodes with

equal frequency, transmissions initiated by the node except the center placed node, will be treated as interference transmissions. Comparing SF with SFS and SFSC, we can see that SFS and SFSC achieve higher throughput. Comparing w/o RTS and w/ RTS, we find that w/ RTS achieves higher throughput at high traffic environment.

Figure 8 shows the number of successfully transmitted and received packets when the average packet generation interval is 0.03 s. In this figure, number of successfully transmitted packets is shown on the left bar, and number of successfully received packets on the right bar. In the right bar, the number of primary, secondary, and synchronous packets, are shown from the bottom to top, respectively.

We can see that w/o RTS has huge number of transmitted packets. However, number of successful received packets significantly decreased. This phenomenon is remarkable especially compared with methods with RTS/CTS exchange. The reason why the number of successful synchronous transmissions behinds the primary and secondary transmissions, is due to collisions with SynT's neighbors. Primary and secondary transmissions send data by exchanging PRTS/PCTS/SCTS packets to prevent collisions, while the synchronous transmission only sends data at a predetermined time.

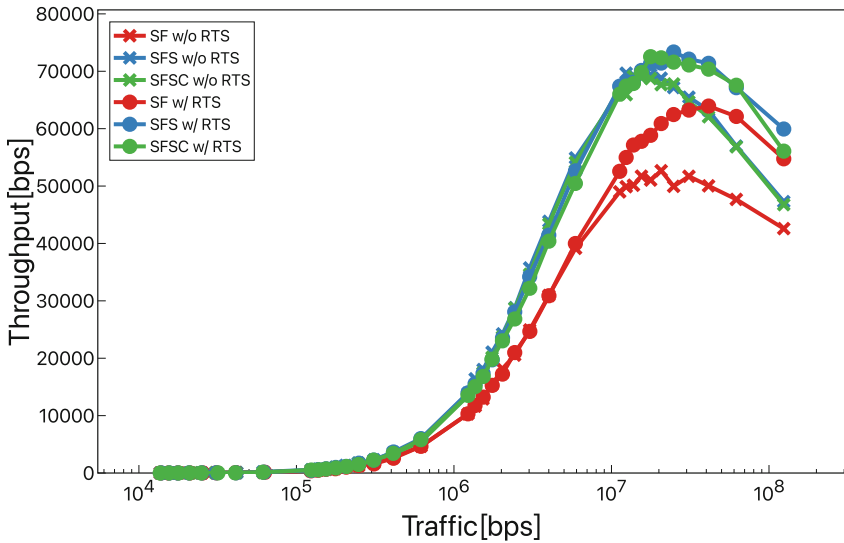


Fig. 7. Throughput performance.

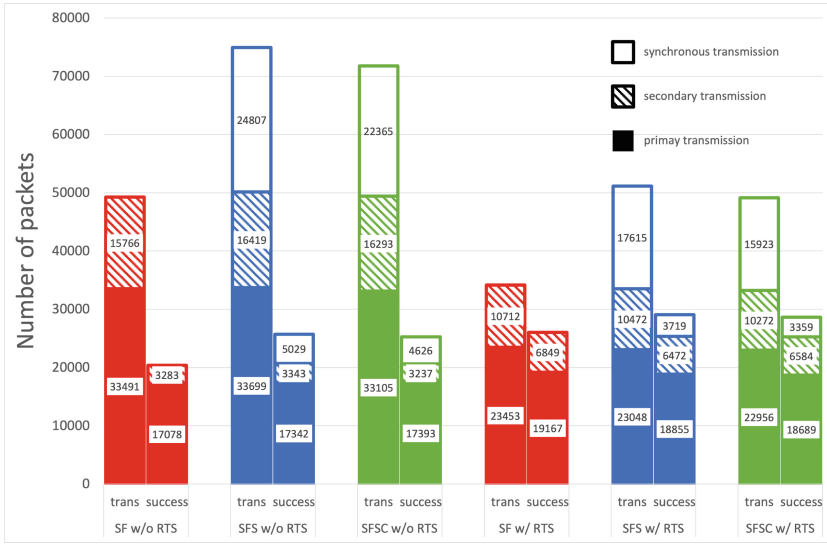


Fig. 8. Number of transmission/reception packets.

4.2 Throughput Characteristics Under Varying Data Size

In this section, we evaluate throughput characteristics while changing Data size. Figure 9 shows the throughput performance when the Data size varying every 512 byte from 512 to 3072. In this figure, all results are evaluated under the condition that the average packet generation interval is 0.05 s.

When the Data size is 512 bytes, there is no large difference on throughput between w/o RTS and w/ RTS of SF, but the throughput of w/o RTS is higher than that of w/ RTS of SFS and SFSC. As the Data size increases, the schemes employing RTS/CTS achieve higher throughput than the scheme without RTS/CTS. This is because the overhead of PRTS/PCTS/SCTS packet exchanged by the proposed method is larger when the Data size is small.

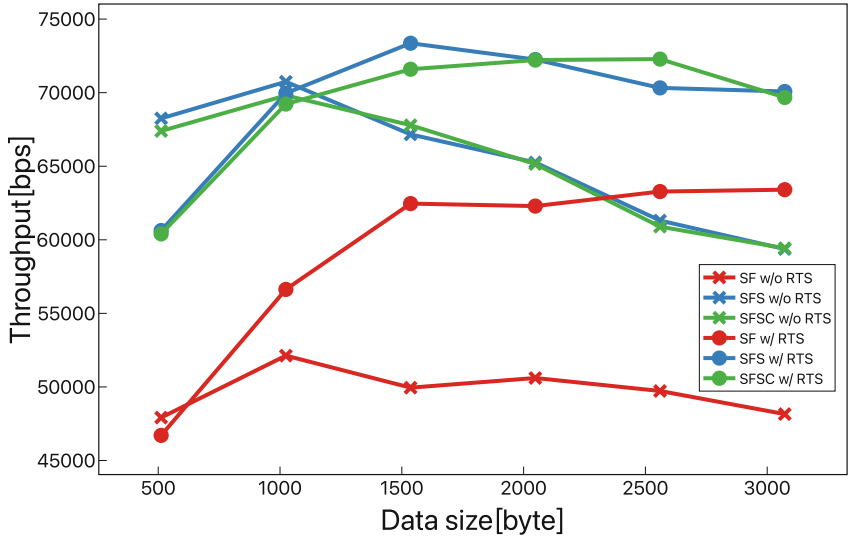


Fig. 9. Throughput characteristics under varying data size.

4.3 Characteristics of Under Varying Number of Nodes Throughput Improvement

The throughput improvement rates when $N = 10, 50, 100, 150,$ and 200 are shown in Fig. 10. In this figure, the average packet generation interval is 0.05 s. Improvement rate is a throughput of each method divided by the throughput of SF w/o RTS.

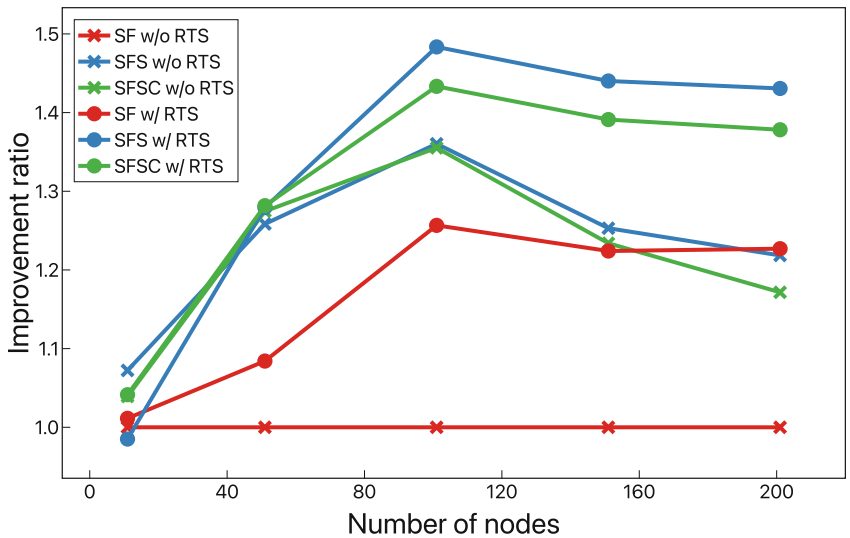


Fig. 10. Throughput improvement rate under varying number of nodes.

As the number of nodes increases, w/ RTS achieves a higher throughput improvement ratio than each method of w/o RTS. Comparing SF with SFS and SFSC, they achieve a maximum throughput improvement ratio is approx. 1.5. Comparing SFS and SFSC, SFS achieve a higher improvement ratio. As the number of nodes increases, the number of neighboring nodes that are candidates for synchronous transmission, also increases. Therefore, the probability of collisions on SynR is low. The case that a lot of nodes are placed in network, increasing synchronous transmission is useful for better performance.

5 Conclusion

For realizing the increase of network capacity for connecting a bunch of IoT devices, this paper, considered a full-duplex wireless communication system that aims to improve throughput performance by capturing transmission opportunities while avoiding collisions. Concretely, we proposed a method of exchanging control packets before Data transmission to suppress new transmissions, and transmitting synchronous transmission by selected PT's neighbors. Our proposal selected candidate synchronous transmitters for avoiding collision between other transmissions. Performance evaluations confirmed that our proposed method well educes throughput performance by increasing synchronous transmission.

References

1. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of Things: a survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* **17**(4), 2347–2376 (2015). <https://doi.org/10.1109/COMST.2015.2444095>
2. Liao, Y., Bian, K., Song, L., Han, Z.: Full-duplex MAC protocol design and analysis. *IEEE Commun. Lett.* **19**(7), 1185–1188 (2015). <https://doi.org/10.1109/LCOMM.2015.2424696>
3. Achaleshwar, S., Gaurav, P., Ashutosh, S.: Pushing the limits of full-duplex: design and realtime implementation. Rice University, Technical report TREE1104 (2011)
4. Jainy, M., et al.: Practical, real-time, full duplex wireless. In: Proceedings of the ACM 17th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2011 (2011)
5. Cheng, W., Zhang, X., Zhang, H.: RTS/FCTS mechanism based full-duplex MAC protocol for wireless networks. In: 2013 IEEE Global Communications Conference (GLOBECOM), pp. 5017–5022 (2013). <https://doi.org/10.1109/GLOCOMW.2013.6855746>
6. Tamaki, K., Ari Raptino, H., Sugiyama, Y., Bandai, M., Saruwatari, S., Watanabe, T.: Full duplex media access control for wireless multi-hop networks. In: 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), pp. 1–5 (2013). <https://doi.org/10.1109/VTCSpring.2013.6692573>
7. Singh, N., Gunawardena, D., Proutiere, A., Radunovi, B., Balan, H.V., Key, P.: Efficient and fair MAC for wireless networks with self-interference cancellation. In: 2011 International Symposium of Modeling and Optimization of Mobile, Ad Hoc, and Wireless Networks, pp. 94–101 (2011). <https://doi.org/10.1109/WIOPT.2011.5930070>
8. Goyal, S., Liu, P., Gurbuz, O., Erkip, E., Panwar, S.: A distributed MAC protocol for full duplex radio. In: 2013 Asilomar Conference on Signals, Systems and Computers, pp. 788–792 (2013). <https://doi.org/10.1109/ACSSC.2013.6810393>

9. Duarte, M., et al.: Design and characterization of a full-duplex multiantenna system for WiFi networks. *IEEE Trans. Veh. Technol.* **63**(3), 1160–1177 (2014). <https://doi.org/10.1109/TVT.2013.2284712>
10. Xie, X., Zhang, X.: Does full-duplex double the capacity of wireless networks? In: *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications 2014*, pp. 253–261 (2014). <https://doi.org/10.1109/INFOCOM.2014.6847946>
11. Duarte, M., Dick, C., Sabharwal, A.: Experiment-driven characterization of full-duplex wireless systems. *IEEE Trans. Wireless Commun.* **11**(12), 4296–4307 (2012). <https://doi.org/10.1109/TWC.2012.102612.111278>
12. Choi, J.I., Jain, M., Srinivasan, K., Levis, P., Katti, S.: Achieving single channel, full duplex wireless communication. In: *Proceedings of 2010 ACM MobiCom*, pp. 1–12 (2010)



Enhancement of Quality Assurance Controls in a Smart Transportation System: Application to Petrol Product Distribution

Rexhina Hoxha^{1,2}, Eva Mandri^{1,2}, Artemisa Sinorukaj^{1,2}, Elinda Kajo Meçe²,
Roberto Sacile¹, Ilir Shinko², and Enrico Zero¹ (✉)

¹ DIBRIS, University of Genova, via Opera Pia 13, Genova, Italy

roberto.sacile@unige.it, enrico.zero@dibris.unige.it

² Faculty of Information Technology, Polytechnic University of Tirana, Mother Theresa Square,
No. 4, Tirana, Albania

{rexhina.hoxha,eva.mandri,artemisa.sinorukaj,emece,
ishinko}@fti.edu.al

Abstract. The goal of this paper is to describe an Intelligent Transport System which manages different kind of entities and sub-systems to ensure safety and security during petrol products transportation. This consists in monitoring in real-time the whole system which makes it possible to enhance its quality control. The data collected from the network allows studying the conformity between the planned quantity and path of distribution with the actual ones. This information is later used for realizing an accurate real-time management system with the aim of identifying possible anomalies. Among the different physical logistics measures which are monitored in real time, temperature has been taken as an important factor which impacts the losses in the quantity of product. In the manuscript, the system design and some results related to specific investigations are reported.

1 Introduction

The process of transporting dangerous goods generates a collection of information which is necessary to be shared between different roles involved in this process. This information is quite compact and sensitive as it may contain the timestamp and the journey of the movement of dangerous goods, details about the storages and other business matters, different trigger messages depending on the incident, etc. At any level, this process should remain transparent, and all this information should be accessible to any person concerned with the development of the process. Furthermore, this data is constantly uploaded, reviewed, and analyzed which can become quite a complex process.

In [1], a system of systems application is described, which provides monitoring of vehicles and detection of anomalies during real-time monitoring of transport in order to provide data about the position of the truck and the quantity of the products loaded and unloaded.

Another monitoring system is described in [2]. The authors propose the use of Radio Frequency Identification (RFID) and Global Positioning System (GPS) technology to

realize the functions of data acquisition in real time, positioning, tracking, and monitoring in transportation. Meanwhile in [3], the authors describe Pecos, a software that can be used to run automatically a series of quality control tests and generate reports which include performance metrics, test results, and graphics.

2 System Architecture

In a daily basis, we are working on improving the Intelligent Transport System controlling and managing the logistics of more than 700 vehicles on the Italian territory. Each vehicle makes two or three tours a day and the mean distance travelled during each tour is about 170 km.

The focus of our management system application is to enhance technical and functional standards with the aim to provide a safe and secure environment for the drivers.

The Intelligent Transport System consist in four main parts:

- On board unit
- Transmission system
- Database
- GIS-based Applications.

2.1 The On-Board Unit

The state of the truck is mainly described by the following information: the position of the vehicle, speed and direction (by GPS and odometer), inclination with respect to the two axes, air suspension pressure, CANBUS information, amount and temperature and pressure of product, state of the load (data comes from the electronic counter) [1].

To gather all the information required to describe the state of the truck in every vehicle, a set of hardware devices and software are embedded in different parts which results in the on-board unit.

On the tractor, the main part is the distribution box which allows the connection to the Controller Area Network, to the odometer, to the emergency button, to the power supply and do the distribution box on the trailer.

On the trailer part different types of devices are found such as: various analog and digital sensors (temperature, pressure, acceleration, inclination with respect to the two axes, and/or status of valves) which detect information to monitor the single parts of the vehicle, the electronic counter, and the concentrator which is equipped with a GPS antenna and a GPRS transmitter/receiver which realize the transmission of real-time data. The concentrator gathers data from all the on-board sources, realizes data processing, includes position information (by a GPS receiver) and makes possible the packing of the data and later it sends these messages on multiple queues at different time intervals. Figure 1 shows the architecture of all these parts and all their connections.

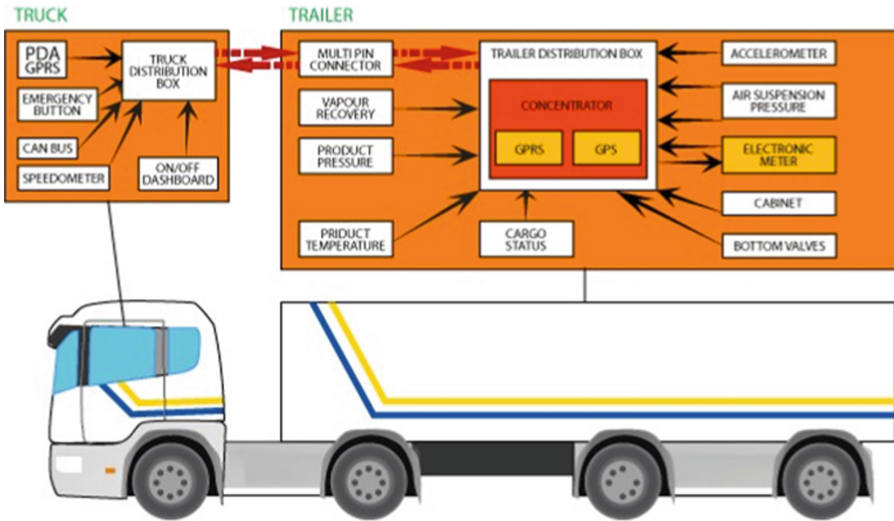


Fig. 1. Architecture of the on-board unit parts

2.2 The Transmission System

Our ITS uses a cluster of centralized servers for receiving and processing data using a Web Service (WS) interface to communicate with the On-board Units, as shown on Fig. 1.

Messages are encapsulated inside a SOAP2 envelope, formatted with XML, and transmitted along with HTTP [4]. Then the messages are sent by a GPRS module equipped with one or more SIM-cards. Between the client and the server, the communication is two-ways since the clients can reply by acknowledging or by more complex commands. After the message is received it is then redirected by the Network Load Balancing system the server which is less overloaded and then is stored in valid and non-valid data in the database.

One of the aspects which is more suitable for our system is that the client/server communication employs a standard data format for the messages that are packed and sent from concentrators. This allows us to have independency between the types of hardware found in the trucks and the higher information layers in our system. What is considered to be a good data format, should accomplish two main conditions: scalability and consistency. Also, it is needed that the rate of the standard-frequency and on-event messages to be distinguished and defined. The first messages give information about telemetry data in a periodical way while the second ones notify a specific on-board event. The approach used consists of a “fixed part” which does not depend on the class of data followed by a “variable part” whose length and format depend on the kind of data. The “fixed part” includes information about creation and transmission date, vehicle identifier, coordinates, and kind of transmitted data [1].

2.3 The Database

The transmission database (T-DB) receives all the raw-messages that come from the Web Service and stores them with the intention of diagnosing them. Due to table locking issues the messages are not parsed in this database. The unpacked messages are moved to the main database server (M-DB) via application server in specific time-intervals. From here data is backed-up in the B-DB which can replace the latter in case of failure.

2.4 Geographical Information System (GIS)

GIS utilizes the hardware, software and all the collected data to manage, analyze and after display in all the possible ways this geographical information. The proposed GIS-based application allows to use effective graphic interface, high scalability, a method to retrieve information from the interface and the ability to perform geographic calculation. Using this application, we can have real-time data about each truck, such as quantity, type of product, position, and the id of the truck. This application assists in data visualization, risk analysis and vehicle routing optimization.

3 Data Collection

The system is also responsible for managing daily deliveries from trucks but also virtual vehicles for simulation purposes, that have a transmission rate equal to a couple of seconds. Throughout this trip the trucks have the possibility to deliver different types of petrol products. The structure of the message that must be sent is made of fields which have data about the geographical position and events which are collected by the sensors in the vehicle. These fields are concatenated in using a semicolon as the separator for two subsequent fields. Thereby, the format of the transmitted message string is:

source_id; transmission_date; reception_date; creation; driver_id; truck_id; trailer_id; CIM; last_MTC; latitude; longitude; data_id; value [5].

This string should be transmitted through GPRS from the real fleet of vehicles and through LAN for the virtual vehicles. The raw messages are then received by the T-DB for diagnostic purposes and from here the unpacked messages are sent in regular time intervals to the M-DB which will be translated then graphically interpreted within the GIS application for visual monitoring.

After this data is stored in our database, we can apply quality check to see if the quantity of product that is distributed has fallen to the same range as the planned one; to check if the trip took the same order as the planned one, and how the temperature of the environment affected the quantity of our product and especially of the LPG.

4 Performance Review

To check the performance of the system, we took data from 2019 (due to pandemic the most recent data is not available). For this year, we check the quantity of product that is distributed in contrast with the planned one. Furthermore, we check for each trip if it has followed the planned path.

4.1 Following the Planned Quantity

As mentioned before, one of the data analysis made is the level of convergence between the quantities that are planned to be distributed and those actually distributed for each depot. We take into the consideration the data from 10 depots with the aim of looking how well the planned quantity has been followed for 2019. The graph presented in the Fig. 2 represents the percentage of deviation in quantity for each depot that we took into consideration. These depots are among those with the largest number of trips during a year.

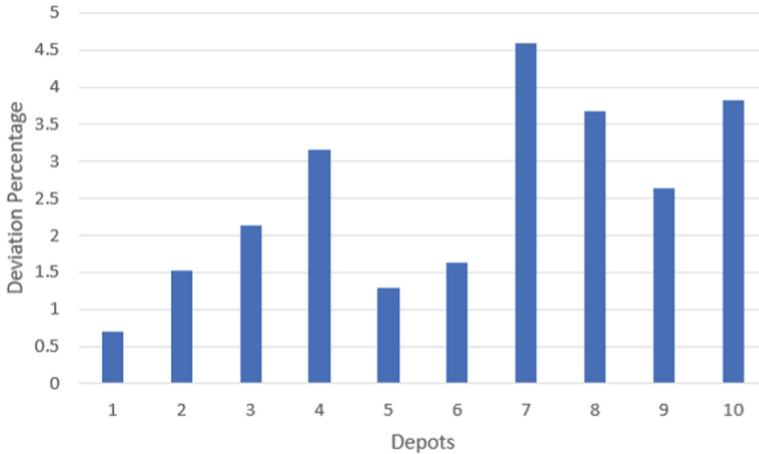


Fig. 2. Percentage of deviation in quantity for 10 depots.

As it can be observed, the maximum percentage of the deviation between the planned and actually distributed quantities is for the depot 7, which results to be also the depot from which the largest number of trips originate.

After this general view of planning vs actual distribution, it is important to deepen the reasons that have led to such deviation. As an example, depot 9 is taken as a case study. We look at all the trips from this depot during the first three months of 2019, in order to identify the tank trucks with potentially large differences between what they had planned to distribute and what they really have distributed.

For instance, in the case of the depot 9, for the first three months of 2019, 88.2% of the trips from this depot had a deviation percentage in the quantities distributed of less than 5%, while the rest of the trips had a deviation percentage of more or equal to 5%.

Further, for the trips that take part in the second category (with a difference percentage greater than or equal to 5%), the tank trucks with the corresponding license plates that have carried out these trips are checked. It is relevant to observe that in the three months studied, there are always the same vehicles (driven in the most of the cases by the same drivers) that, out of the total number of trips made, have a high percentage of second category trips. This leads us to the detection of one of the possible causes of the problem, a failure in performing the tasks properly by the responsible person in our case the driver.

Thus, for instance, if we look at the percentages of the difference between the planned values and those actually distributed for each type of different product distributed by the tank truck, the percentages of difference result to be very small for specific types of products, only that in a few cases the driver has failed to set the exact value that was downloaded, which makes the total percentage increase (e.g. for a certain planned value of *Gasoline*, the downloaded value results to be set zero).

In other cases, it can be seen that regardless of the type of product distributed, the total distributed quantities are very close to the planned ones. However, the deviation percentages for each type of product, result to be large values, what means that the driver has failed to set the right type of product that he has downloaded at a certain moment.

4.2 Following the Path

To check if the trucks are following the planned path and correct sequence of deliveries to service stations, we take data for the actual trip that happened and for the plan of this trip. We take into consideration data from 10 depots, and we have also unclassified data with respect to the depot due to errors on transmission or no upload event found.

In the figure below, we can see the percentage of deviance of trips from the planned ones (Fig. 3).

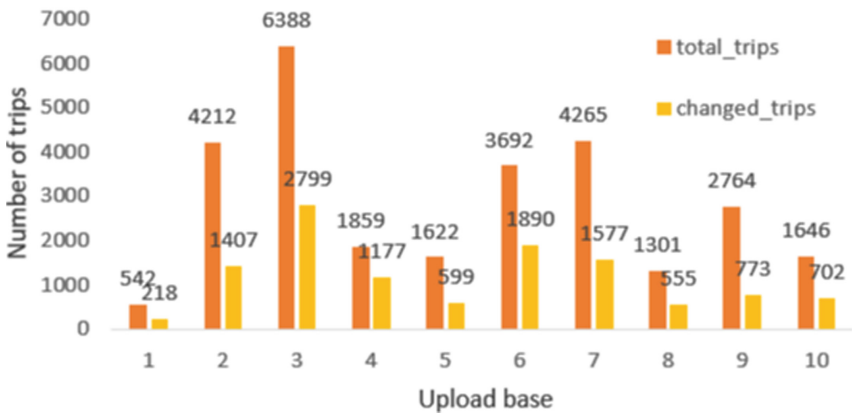


Fig. 3. Total number of trips and the changed trips for 10 depots

It is noticeable that the amount of trips which do not follow the planned sequence is high. For 2019, about 50% of the trips had at least one change on the stations order. This is not quite favorable for us, because in a other investigations, we take the plan as the correct path by default. In order to increase the performance of the system, we should take this factor into account when designing other solutions and when we create new services for the system. One feature that is required by the company is to be notified for each next stop when the truck finishes the previous delivery. Due to unknown reasons for now, the order of stops can be different. This means that we are left unknown for the exact future station where this truck is going. This poses a lot of risk factor and difficulty in managing the trip.

It is also possible for many trips not to visit all the planned stations where they had to deliver the dangerous goods. For 2019 we took the same base stations as before and the result is as in the figure below where 16.4% of the trips were not finished which means that there exists a problem in why the process was not concluded (Fig. 4).

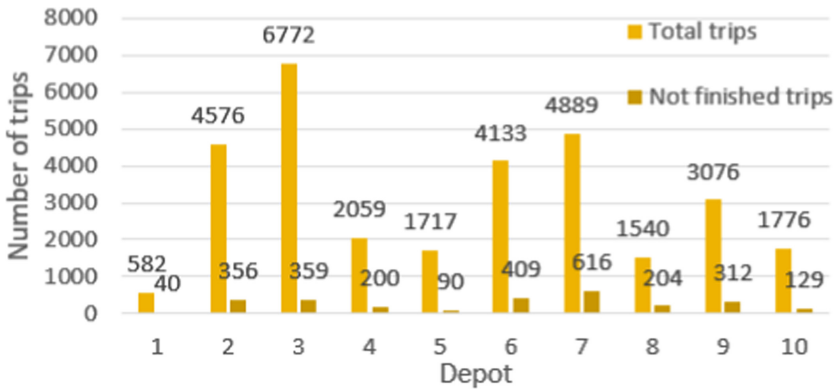


Fig. 4. Total number of trips and the not-finished trips for 10 upload bases

4.3 The Role of Temperature While Transporting LPG

Liquefied petroleum gas (LPG) is one of the dangerous goods transported by this company. LPG is a fuel gas made of petrol which contains a flammable mixture of hydrocarbon gases mostly propane (C₃H₈) and butane (C₄H₁₀). Therefore because of his nature, LPG needs a different approach for storing and transporting. This is conducted by tank trucks. Along with all the factors that impact (effect) the product loss for LPG, temperature is an important one. Just like other gases, LPG will react the same, the volume will expand inside the truck if the temperature increases and the volume will shrink inside the truck if the temperature decreases. Considering this information, we expect to have differences in levels of the gas throughout the journey. Because of all these level differences we should keep in mind the risks that could bring. In order to keep a safe travel, the tank truck will be filled around 80%--85% of the whole capacity so we can leave free space for the differences of the level throughout the journey.

5 Implementation of the Management System

For each day, we have an average of 750 trips, and for each one of these trips we have an average of 3 stops with a maximum of 6 types of fuels possible to distribute. For a company to measure the quality of service it is necessary to compare the convergence between the planned quantity and the actual distributed quantity of fuel. LPG represents a specific interest in our work because it has different way of distribution, and it is more difficult to be managed.

For our system to deliver real-time data we should be able to know with high accuracy the details of the ongoing trip. It would be quite easier to manage the system if the planned quantity or the order of stops of stations would be followed. To make it easier for the end user to check different aspects of the ongoing or other passed trips TIP offers webpages which are developed in VB.NET and the data is filtered and selected from the servers using Stored Procedures.

5.1 Quantity Distributed

Having a service that allows the user to view real-time information on the status of the quantities of distributed products is important to ensure a safe operation of the system by detecting possible anomalies and taking measures to correct them. This is precisely the purpose of the created web page. The user can check in real time the distributed quantities against those planned based on the filters he requires.

Fig. 5. The research form of the quantity check Web Page.

Figure 5 shows the research form of the web page added to the system, where as it can be seen, a user can have information in real time as well as historical data based on the selected values of the depot, supplier, license plate, time period as well as the percentage of deviation in quantity.

For the selected values of the filters, the information that the user receives is of different forms:

- Graphs that show for each day within the selected time interval the planned values against those distributed for each type of product. The data can be presented for any selected depot, supplier, license plate or a combination of the three filters.

- Histograms which, for a selected depot, a time period and an interval of percentage deviation in quantity, present the license plates of tank trucks with respect to the frequency of deviation in terms of trips. The frequency of deviation represents the number of trips made by the respective vehicles with deviation percentage within the selected interval with respect to the total number of trips made by this vehicle.
- Tables with detailed data on vehicle license plate, date, planned quantities, real distributed quantities, difference and type of product.

5.2 Following the Planned Path

Knowing the status of the trip at each chosen time period is a key factor in a real-time data management system. The first thing that is done to firstly store the messages in the database is to identify the trip and the event of which it belongs. It is a work on development to make the service of trip identification online, because until now we get the information from the messages the day after they happen. With this characteristic in mind, it will also be possible to manage and check different aspects of the trip while it is ongoing. The user can see the service station where the truck will be, the quantity that it has distributed, if the path has changed and where the other previous stops were. To see the behavior of the truck on time we create a webpage so the user can see if there are anomalies or not acceptable results, and it can react and manage to find a solution. From this webpage the user can also check previous trips if they want to gather statistics on a specific upload base or year. The data about the trip is shown as in Fig. 6.

Soste Eseguito	Codice NSI	Nome di Base di Carico	Codice ENI	Targa Posteriore	Targa Anteriore	Dettaglio
6/6	4566287001	SANNAZZARO	456	XA605JN	FF564ER	
3/3	4566288006	SANNAZZARO	456	XA605JN	FF564ER	
3/3	4566288007	SANNAZZARO	456	XA605JN	FF564ER	

Fig. 6. Web-page representation for following the details of the ongoing trip

We can also have in real time notifications for every event and also for the events to come. One solution would be to utilize the location to check for the proximity of the truck to the planned stations and when we are within the allowed threshold then we can update the status of the trip.

Let’s assume that the trip shown in the Fig. 7 is a planned trip which is meant to be done.

The planned trip has the order of stations in a chronological one, but the truck driver decides to do 5-4-3-1-2. If we were to base the prediction of where the truck is going to be only by the planned trip then we would be mistaken, because we would say that the second station would be number 2 and the first one was number 1. To implement a more correct solution we take into consideration the location of the truck for each telemetry message that we receive. If the truck is inside the red circle of each station, then we would be sure that the next stop would correspond to the correct one. This is a new service that is being implemented for TIP in order to take the anomalies that were explained into account so we can raise the efficiency of the whole system.

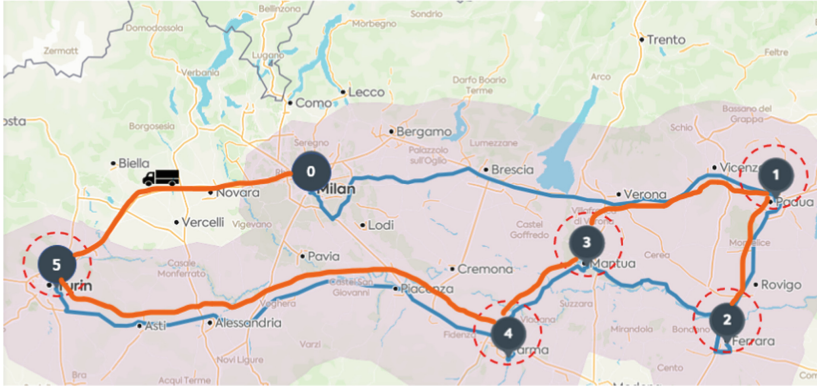


Fig. 7. An assumed trip with the order of the stations planned to follow and the real order followed

5.3 Calculating the Difference Between Levels Relating to Temperature

As we mention before, throughout the journey we are expecting differences in the level of the tank truck. In the database we have implemented a formula that takes into account how the difference of temperatures affects LPG while it's on the tank. The formula is as below:

$$V = V0 + V0 * 0.002 * \delta T$$

- V - is the current volume of the gas
- $V0$ - is the initial quantity of the gas
- δT - is the difference between the temperatures

In order to see if any differences exist, we need the information below:

- the volume when the truck enters the service station
- the volume when the truck exits the service station
- the temperature at every service station
- the quantity downloaded at every service station

We can take as an example a trip which has to deliver LPG in three different service station.

- First service station: Firstly, we calculate the volume with the formula when the truck arrives at the service station. The initial quantity is the whole quantity uploaded at the base and the temperature will be the difference between the temperature at the base and the temperature at the current service station. The volume that the truck exits the service station will be the difference between volume before it enters and the quantity downloaded at the service station.
- Second service station: Even here we will do the same calculation as before but now the initial quantity will be the quantity that the truck has after it exit service station

one and the temperature will be the difference between temperature at service station one with the temperature at the current service station. In the same way as before we will calculate the volume of the truck after it exit the service station

- Third service station: We will follow the same way as we did in service station two.

After all the calculations have been made. We will be able to see if there are any differences in volume or not. It will be shown that only in the journeys where the temperature was constant there were no differences in the levels of the volume, in every other case we would have more or less product depending on the temperature.

6 Conclusions

We have shown that there exist anomalies in our monitored logistic system which should be taken care in order for us to enhance the best quality of service.

Taking into consideration the fact that half of the trips changes the order of stations helped us into simulating the other proposed method explained in Fig. 7. Before taking into consideration the anomaly, we made analysis which were wrong 50% of the time. When we simulate the new method, we decreases mistakes to 30%. We get a 20% raise in performance quality. This is quite a valuable enhancement for a real-time working system.

Moreover, the quality control on the quantity of distributed products, led us to one of the potential reasons for the anomalies that existed in the proper following of the planned quantities, that is the driver not performing the task properly.

Lastly, due to different levels, we would have throughout the journey, we came into the conclusion that temperature was indeed a key factor while transporting LPG. We saw how the differences of temperatures affected the level of LPG during the travel.

References

1. Benza, M., et al.: Intelligent transport systems (ITS) applications on dangerous good transport on road in Italy (2012)
2. Miao Yu, J.F., Deng, T.: Application of RFID and GPS technology in transportation vehicles monitoring system for dangerous goods (2012)
3. Klise, K.A., Stein, J.S.: Automated performance monitoring for PV systems using Pecos (2016)
4. Gudgin, M., et al.: W3C Recommendation, SOAP Version 1.2 Part 1: Messaging Framework, 2nd edn. World Wide Web Consortium (W3C), April 2007
5. Laarabi, M., Boulmakoul, A., Sacile, R., Garbolino, E.: A scalable communication middleware for real-time data collection of dangerous goods vehicle activities. *Transp. Res. Part C Emerging Technol.* **48**, 404–417 (2014). <https://doi.org/10.1016/j.trc.2014.09.006>. Accessed 5 Aug 2022



Hardware-Software Interworking Real-Time V2X Dynamic Analysis Method

Insu Oh¹, Munkhdelgerekh Batzorig², Baasantogtokh Duulga¹, and Kangbin Yim¹ (✉)

¹ Department of Information Security Engineering, Soonchunhyang University, Asan, Korea
{catalyst32, prab11, yim}@sch.ac.kr

² Department of Smart Convergence Security, Soonchunhyang University, Asan, Korea
munkhdelgerekh@sch.ac.kr

Abstract. V2X communication technology, which is autonomous and cooperative driving, is developing, and various V2X products for application to vehicles are being developed. However, the test environment for testing this is not sufficient. Also, security threats targeting V2X are increasing, and most of the test frameworks to verify them depend on the simulation environment. Therefore, in this paper, a V2X communication test environment is constructed using hardware and software, and a dynamic analysis method is described by collecting V2X datasets in real time.

1 Introduction

With the recent development of self-driving technology, self-driving cars are being developed to identify external situations through external sensors such as cameras and radars for safety and convenience, and accordingly, driver interference is minimized and dependence on cars has increased. For the safety of drivers, the law mandates that autonomous driving technology is essential for recently released cars. However, if the sensor fails to operate normally due to errors and attacks for autonomous driving, huge accidents can occur, and autonomous cooperative driving is required for this.

The communication technology of autonomous cooperative driving is largely divided into Vehicle to Everything (V2X), Vehicle to Infrastructure (V2I), Vehicle to Vehicle (V2V), and Vehicle to Nomadic Device (V2N) that communicates and exchanges information with neighboring entities through networks connected to moving vehicles. It provides telematics services, automatic fare collection services, and traffic information collection and provision services using V2I networking with vehicle and road infrastructure and communication functions. The V2V network may provide a service through communication between safe vehicles and provide cooperative driving services by delivering real-time vehicle information [1].

In particular, the V2N network may provide a vehicle diagnosis and control service by directly connecting the portable terminal and the vehicle. This increases the likelihood of an attack in which an attacker can access a car. Using this approach, an attacker can harm human life through malicious behavior on a vehicle, so a reliable review and verification study is needed before applying these external communication devices to a vehicle.

With the recent increase in interest in V2X security, research on the development of the V2X framework is underway, but it is far from the actual environment to analyze security technologies that will be applied in the future. There is a lack of experimental environment for security technology, and safety is low to apply the security technology used in the simulation environment to actual vehicles. Therefore, in this paper, based on the V2X technology to be applied in real vehicles in the future, we propose a simulation environment that can be tested in various ways through a V2X communication environment consisting of real equipment such as OBU and RSU.

2 Related Works

Although recently released vehicles are equipped with V2X equipment, V2X technology can be applied to existing vehicles through additional equipment such as OBU. When V2X is commercialized and widely used, there is a possibility that various security threats will occur. Therefore, in order to verify the safety from these security threats, it is important to configure the V2X research environment for testing, develop a framework for communication testing, and develop scenarios applicable in the real environment. However, most of the simulation environment is a framework developed without considering the actual situation, and a method to supplement this is needed.

Various dynamic studies are being conducted to apply security technology to the V2X system that is currently being developed, and recently, Veins Framework is frequently used. Veins Framework is a simulation environment that configures V2X networks and provides data analysis functions. V2X simulation is largely divided into simulation test framework and hybrid test framework [2].

2.1 V2X Security Threats

V2X communication requires security for authentication of one's identity, reliability of transmitted messages, and non-repudiation of message reception due to the characteristics of wireless communication. In a situation where various types of objects communicate with each other through wireless messages, various security threat problems may occur, and DoS and Sybil Attacks False data injection may occur typically [3, 4].

Denial of Service (DoS) Attack. By terminating or stopping the network established by the RSU, communication control between vehicles is forcibly stopped. This makes it difficult to detect as the attacking node launches the attack from a different location. Since V2X communication is wireless communication, an attacker at the physical layer can limit the transmission/reception of messages by interrupting the communication channel through signal interference. This may reduce the network reliability of the V2X system.

Sybil Attack. Possible by creating fake identities that do not exist and sending messages. The attacker's vehicle can use several fake identities to appear as other moving vehicles that are not real, to make the road look congested, or to transmit misleading information about road conditions to nearby vehicles or RSUs. It also destroys the identity of a specific vehicle, which can cause confusion and inconvenience to normal users.

False Data Injection Attack. As an attack that injects false data, normal users can receive incorrect information due to not only location and sensor data, but also object and object information. It can have a fatal impact on platooning services.

2.2 V2X Simulation Test Framework

The V2X simulation test framework is mainly used in V2X communication theory research, and the virtual test framework consists of a network simulator, a traffic simulator, and an application simulator. We exchange information from three simulators and form a virtual simulation environment through interaction [5].

Network Simulator. Simulators such as NS2, NS3, OMNet++ and OPNet exist and are mainly used to generate V2X protocols for simulating communication between vehicles and road devices and to configure dynamic topology of nodes (vehicle, pedestrian, etc.) [6].

Traffic Simulator. It is used to build traffic and vehicle models or to generate relevant examples. There are existing SUMO, VISSM, TransModeler, CORSIM, etc. [7].

Application Simulator. It is used to create various computing applications through programming languages to construct scenarios. Currently, the most representative virtual test frameworks are TraNS and Veins [8].

2.3 V2X Hybrid Test Framework

Virtual test frameworks generally use virtual environments based on virtual data and are difficult to reflect real situations. However, there is a risk to real environment test, so a hybrid test method that combines real and virtual is needed.

Hybrid test frameworks include Software-in-the-loop (SIL), Hardware-in-the-loop (HIL), and Vehicle—in-the-loop (VEHIL) [9]. In the same framework, we combine HIL and SIL tests and propose an intelligent vehicle hybrid simulation tool, Virtual Intelligent Vehicle Urban Simulator (VIVUS) [10]. In addition, we propose a parallel test based on virtual-real-world interactions in autonomous vehicle competitions [11]. Based on this, we propose a cloud-based Cyber-Physical-Social System (CPSS) that combines Cyber-Physical Systems (CPS) and parallel operation [12].

3 V2X Communication Message Structure

V2X communication is largely divided into C-V2X (Cellular Vehicle to Everythings) using LTE and DSRC based on WAVE (Wireless Access Vehicular Environments) protocol. Most V2X systems utilize DSRC, but the current trend is to move to C-V2X due to cost issues for installing additional antennas. In addition, messages are defined through the standard for communication between vehicle and vehicle communication vehicle and infrastructure, which is summarized in SAE J1735. Therefore, the basic contents of V2X communication and the structure of data to be collected will be described.

3.1 DSRC and C-V2X

DSRC is a short-distance wireless communication method dedicated to ITS (Intelligent Transport System) and is being used as a communication module for automatic road toll collection systems worldwide [13]. As part of short-range wireless communication, various types of information are transmitted and received using the 5.8 Ghz frequency band. WAVE communication is a wireless LAN-based technology specialized for high-speed mobile environments defined by IEEE 802.11p. With the addition of high-speed movement and large-capacity data transmission functions, it is necessary to construct many facilities for each communication radius with a short coverage of about 500 m. Lastly, C-V2X is a cellular-based V2X system specified in 3GPP Rel.14, and in Europe, ETSI has defined EN to utilize C-V2X as an access layer technology for ITS. Because it is based on mobile communication, only the amount used is charged [14].

3.2 SAE J2735 Message Format

The format and structure of messages, data frames, and data elements for data exchange between V2V and V2I are defined, and various types of messages exist [15].

BSM (Basic Safety Message). In V2V, a vehicle-to-vehicle communication, a message related to safety provides situational data (position, direction, speed) used to evaluate the threat potential to surrounding vehicles. Representative features include Forwarded Collision Warning (FCW), Emergency Electronic Brake Lights (EEBL), Do Not Pass Warning (DNPW), and Left Turn Assist (LTA).

RSA (Road Side Alert). It is a message used for communication between the roadside base station and the vehicle, and includes information such as road danger conditions, and is a message used when the roadside base station transmits road danger conditions to nearby vehicle terminals.

SPaT (Signal Phase and Time). It is an information transmission message according to the traffic lights of the intersection, and the current signal status and expected change time for each lane are provided by the RSU connected to the traffic lights. It is sent from the RSU to provide geographic information of intersections and links the status of SPaT with geographic information.

4 Real-Time V2X Datasets Collection Methods

In order to build an actual V2X system, OBU/RSU devices for V2X communication by connecting with a vehicle are required. Therefore, the simulation environment was constructed using commercial equipment currently on sale. Company A's V2X system, which supports both C-V2X and V2X systems, can be configured to meet U.S. and European standards according to the desired settings. DSRC standard messages can be transmitted according to SAE J2735, which defines structures such as messages for data exchange between V2V and V2I, data frames, and data elements and formats. Since all messages in this V2X are transmitted wirelessly, security is required because all nodes around it can be received and interpreted. Representative security factors include data encryption and certificates to enhance the security of V2X.

4.1 V2X OBU/RSU Equipment Environment

To establish an autonomous cooperative driving V2X communication environment, OBU terminals were used to connect to real vehicles, and RSU antenna devices installed on roads were installed to design a demonstration system that can transmit and receive DSRC messages through the V2X IEEE 1609.2 WAVE protocol. And implemented. A real simulation environment capable of V2V and V2I communication was installed on the campus (Figs. 1 and 2).

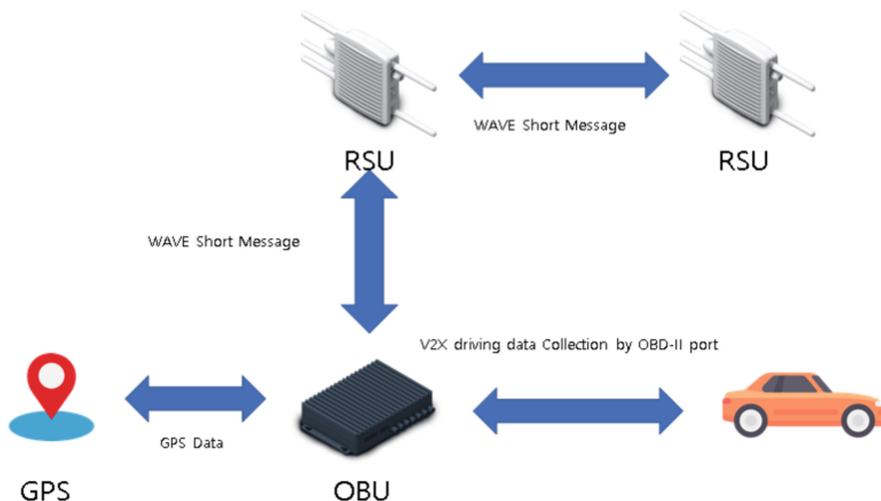


Fig. 1. Hardware-software interworking real-time V2X data collection diagram

OBU (On-Board Unit). Consists of a terminal, GPS, an antenna capable of transmitting 4.9 Ghz, and a display that delivers information to the driver. It transmits vehicle information to the outside through messages in real time and receives surrounding situations through messages. The same environment was installed in two different vehicles to form a V2V environment. In addition, OBU analyzes the vehicle's internal network and transmits messages such as vehicle status and location information through DSRC.

RSU (Road-Side Unit). It is a device that controls network messages used in V2X communication, and mainly serves to deliver OBU network messages within the RSU range. It is also used for services related to traffic lights and vehicles in connection with infrastructure. RSU devices were installed along the road at regular intervals so that vehicles could communicate with each other on the school campus.



Fig. 2. Actual OBU hardware environment installed on Car

4.2 V2X Dataset Collection

Since communication between OBU/RSU devices is based on standard Ethernet IEEE1609, packets transmitted through V2X communication can be monitored. If you check the message through Wireshark, you can understand the structure of the V2X message packet, and you can check the packet containing the most basic information, the GPS information of the OBU (Fig. 3).

V2X communication uses DSRC technology to transmit various vehicle information from the OBU device to the outside, including in-vehicle network information (STI) and vehicle location information (NAV). The transmitted message is transmitted by composing a message based on the SAEJ2735 standard, which defines the format and structure of messages, data frames, and data elements for data exchange between V2V and V2I.

The BSM (Basic Safety Message) message is the most basic message and is a message related to safety in V2V communication between vehicles. It provides surrounding vehicles with location, direction, and speed information, which are situational data used to evaluate threat potential. Message information includes Forwarded Collision Warning (FCW), Emergency Electronic Brake Lights (EEBL), Do Not Pass Warning (DNPW) and Left Turn Assist (LTA) functions.

```

▼ nav: Navigation
  valid: true
  gpsTimestamp: 2021- 11- 25 21:35:38.000 +09:00
  latitude: 36.7691217 degree
  longitude: 126.9316097 degree
  altitude: 104.98 m
  heading: 360.0 degree
  speed: 0 m/s (standstill)
  posConfEllipseSemiMajor: (<=) 2.00 m
  posConfEllipseSemiMinor: (<=) 2.00 m
  posConfEllipseSemiMajorOrientation: 45.0 degree

```

```

00 00 00 00 00 00 00 00 00 00 00 00 08 00 45 00
00 4A 00 00 40 00 40 11 2C 4E C0 A8 01 36 00 00
00 00 75 30 1F 07 00 36 00 00 13 21 AC 20 1A 00
00 00 03 01 00 00 01 7D 57 18 1D 90 15 EA 85 D1
4B A8 3A 01 00 00 29 02 0E 10 00 00 00 C8 00 C8
01 C2 01 2C 00 7F 00 2D

```

Fig. 3. Actual collected datasets containing GPS information and V2X messages include latitude and longitude

4.3 Real Time V2X Dataset Collection

The structure of the method of collecting data in real time through the previously built V2X system simulation environment is shown in the figure above (Fig. 4).

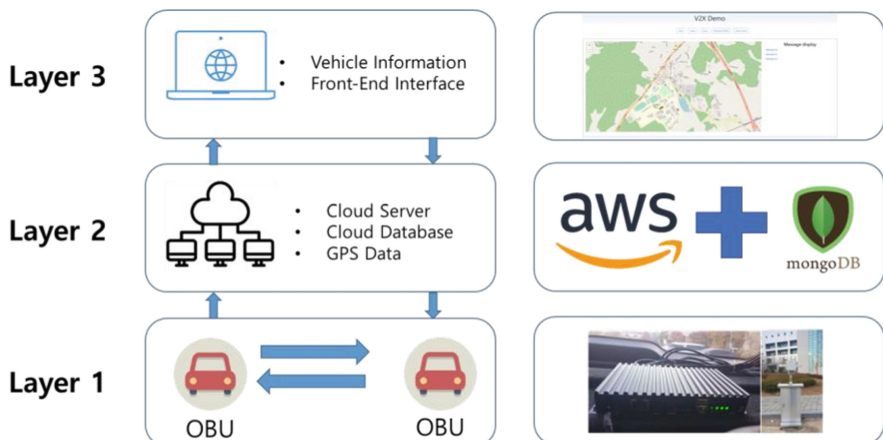


Fig. 4. V2X simulation framework

Data Collection Layer. Before collecting V2X-related data, follow the scenario to determine the data you want to extract. You can check the C2P messages communicated by the OBU connected in real time through a packet analysis tool by connecting to

the OBU device connected to the vehicle via Wi-Fi. In this way, in Layer 1, the vehicle is driven according to the scenario and the collected data is saved in “.pcap” format. It plays the role of uploading the saved file to the cloud server.

Cloud Server Data Analysis Layer. In Layer 2, a cloud server was configured in Amazon Server (AWS) for cloud computing. You can check the position and velocity data by following the C2P protocol for incoming messages to the server. Location information starts from the 52nd byte of the C2P message. In the location information, you can check the data of the time the message was sent, latitude and longitude, and speed. These data are calculated in byte steps in the cloud and stored in the database.

Data Visualization Layer. Finally, in Layer 3, HMI (user interface) was designed and a server was built to visualize data. The website displays the data collected while moving at the speed of the vehicle based on the vehicle’s location data as shown in the figure below.

5 Conclusion

V2X stands for communication between the vehicle and everything else. These advances have broadened the attack range that attackers can access. Security technology must be applied to the vehicle as an attacker can directly and remotely control the vehicle with this access and cause physical damage to the vehicle, environment and people. However, it is difficult to study these security techniques and apply them to real situations.

Therefore, in this paper, we propose a safe experimental framework for V2X communication constructed based on real data. This framework is a cloud environment that anyone can use by utilizing RSU and OBU devices mainly used in V2X communication. Using this, it is thought that analysis in a simulation environment through V2X data will be easy.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A2001810) and This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 2022-0-01197-0, Convergence security core talent training business (SoonChunHyang University)).

References

1. Song, Y.-S., Jo, H.-B.: V2X 통신 기술 및 서비스 동향. *Commun. Korean Inst. Inf. Sci. Eng.* **31**(1), 19–24 (2013)
2. Sommer, C., German, R., Dressler, F.: Bidirectionally coupled network and road traffic simulation for improved IVC analysis. *IEEE Trans. Mob. Comput. (TMC)* **10**(1), 3–15 (2011)
3. Hasan, M., Mohan, S., Shimizu, T., Lu, H.: Securing vehicle-to-everything (V2X) communication platforms. *IEEE Trans. Intell. Veh.* **5**(4), 693–713 (2020)

4. Oh, I., Jeong, E., Park, J., Jeong, T., Park, J., Yim, K.: Cyber attack scenarios in cooperative automated driving. In: Barolli, L., Takizawa, M., Enokido, T., Chen, H.C., Matsuo, K. (eds.) *Advances on Broad-Band Wireless Computing, Communication and Applications*, vol. 159, pp. 416–425. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61108-8_41
5. Choudhury, A., Maszczyk, T., Math, C.B., Li, H., Dauwels, J.: An integrated simulation environment for testing V2X protocols and applications. *Procedia Comput. Sci.* **80**, 2042–2052 (2016)
6. Siraj, S., Gupta, A., Badgujar, R.: Network simulation tools survey. *Int. J. Adv. Res. Comput. Commun. Eng.* **1**(4), 199–206 (2012)
7. Saidallah, M., El Fergougui, A., Elaloui, A.E.: A comparative study of urban road traffic simulators. In: *EDP Sciences*, p. 05002 (2016)
8. Sommer, C., et al.: Veins: the open source vehicular network simulation framework. In: *Virdis, A., Kirsche, M. (eds.) Recent Advances in Network Simulation*, pp. 215–252. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12842-5_6
9. Tettamanti, T., Szalai, M., Vass, S., Tihanyi, V.: Vehicle-in-the-loop test environment for autonomous driving with microscopic traffic simulation, pp. 1–6. *IEEE* (2018)
10. Gechter, F., Dafflon, B., Gruer, P., Koukam, A.: Towards a hybrid real/virtual simulation of autonomous vehicles for critical scenarios (2014)
11. Li, L., et al.: Parallel testing of vehicle intelligence via virtual-real interaction. *Sci. Robot.* **4**(28), eaaw4106 (2019)
12. Wang, F.-Y., Zheng, N.-N., Cao, D., Martinez, C.M., Li, L., Liu, T.: Parallel driving in CPSS: a unified approach for transport automation and vehicle intelligence. *IEEE/CAA J. Autom. Sin.* **4**(4), 577–587 (2017)
13. Kenney, J.B.: Dedicated short-range communications (DSRC) standards in the United States. *Proc. IEEE* **99**(7), 1162–1182 (2011)
14. Garcia-Roger, D., González, E.E., Martín-Sacristán, D., Monserrat, J.F.: V2X support in 3GPP specifications: from 4G to 5G and beyond. *IEEE Access* **8**, 190946–190963 (2020)
15. Hedges, C., Perry, F.: Overview and use of SAE J2735 message sets for commercial vehicles. No. 2008-01-2650. *SAE Technical Paper* (2008)



Location-Based Autonomous Transmission Control Method for Spatio-Temporal Data Retention System

Daiki Nobayashi¹(✉), Kazuya Tsukamoto¹, Takeshi Ikenaga¹, and Myung Lee²

¹ Kyushu Institute of Technology, Fukuoka, Japan

{nova, ike}@ecs.kyutech.ac.jp, tsukamoto@csn.kyutech.ac.jp

² City University of New York, City College, New York, USA

mlee@ccny.cuny.edu

Abstract. With the development and spread of IoT technology, various devices have been connected to networks. Some data generated from IoT devices depends on geographical location and time (Spatio-Temporal Data). The realization of an architecture for “local production and consumption of STDs” can contribute to location-dependent applications, and therefore we have proposed the STD retention system with vehicles. In our previous study, the vehicle controlled the data transmission probability according to the density of the neighboring vehicles in order to reduce the data transmissions. However, since this method requires all vehicles to transmit beacons, it suffers from the excessive beacon collision when the vehicle density becomes high. In this paper, we propose a data transmission control method that realizes STD retention without transmitting beacons. Our simulation results using Luxembourg model demonstrates that the proposed method can achieve high coverage rate while decreasing the number of data transmissions compared with the existing transmission method based on transmission probability control in real environments.

1 Introduction

With the development and spread of the Internet of Things (IoT) technology, various devices, such as personal computers, smartphones, home appliances, automobiles, various types of sensors, and so on, have connected to networks. According to the Cisco Annual Internet Report (2018–2023) [1], the number of Machine-to-Machine (M2M) devices, which are typical examples of IoT devices, is expected to be approximately 14.7 billion by 2023 and further increases are expected in the future. In today’s IoT services, data generated from IoT devices is collected and analyzed on a cloud server via the Internet. Therefore, as the number of IoT devices increases, the traffic on the backhaul network also increases, so the load on the existing infrastructure increases.

On the other hand, data generated from IoT devices include traffic information, weather information, disaster information, temporary store advertisements, and so on, depending on the location and time of data generation. We have defined such data as spatio-temporal data (STD). The STD pertains to the concept of “local production and consumption,” useful to local users in the vicinity of which it is generated. For example, in the case of a traffic accident, the accident information from the vehicles involved in the accident disseminates to surrounding vehicles so that the drivers in the vicinity of accident location passively obtain traffic accident information without having to access the cloud. Finally, the drivers take action to avoid road closures and traffic jams.

To achieve a novel network architecture for the distribution of STD, we have proposed the STD retention system (STD-RS) using vehicles [2]. In the STD-RS, information hubs (InfoHubs) are defined as vehicles that forward STDs and provide STDs directly to users. By broadcasting the STD, InfoHubs not only spread and maintain the data in a specific space, but users can passively obtain the STD. However, as the number of InfoHubs transmitting data increases, data collision also increases, making it difficult to achieve the objective of STD-RS. It is important to minimize data transmission in order for InfoHubs to effectively retain STDs. Therefore, we proposed a transmission control method based on the number of neighboring vehicles [2–5]. In these methods, each vehicle broadcasts a beacon periodically to notify its presence to neighboring vehicles. Thus, each vehicle estimates the number of neighboring vehicles from the number of received beacons and measures the number of received STD from neighboring vehicles. In addition, the number of data transmissions was reduced by setting the data transmission probability based on the estimated number of neighboring vehicles and the number of received data for each data transmission interval. However, as the number of vehicles increases, the number of beacon transmissions also increases. As a result, the collision among data and beacon transmission increases, severely wasting wireless resources.

In this paper, we propose the location-based autonomous transmission control method without beacons for the STD-RS. This method defines the appropriate positions for data¹ transmission in a specific area based on data transmission distance. Each vehicle controls data transmission according to the relationship among its defined position, its own position, and the data transmission status of neighboring vehicles. To verify the feasibility and the effectiveness of the proposed data transmission control scheme in a practical scenario, the Luxembourg model (LuST) is used for our performance study, which simulates 24-h vehicular traffic in Luxembourg City, as a traffic model.

The remainder of this paper is organized as follows. In Sect. 2, we discuss related works for the STD-RS. In Sect. 3, we describe the STD-RS, and then Sect. 4 describes our proposed method. Section 5 provides the simulation model, results, and discussion. Finally, we provide conclusions in Sect. 6.

¹ We also simply describe the STD as data when explaining the proposed method.

2 Related Works

The geocast technology, which transmits messages to predefined geographic areas, is expected to enable location-dependent applications and services. Various geocast routing protocols are introduced in [6]. Among them, Abiding Geocast is a mechanism to deliver messages periodically within the lifetime of the message. Literature [7] proposes three component approaches to achieve Abiding Geocast using vehicles: server, election, and neighbor approaches. In server and election approaches, specific servers and vehicles must collect other vehicles' location information and deliver messages based on geocast routing. Therefore, the load on the server and the vehicle is large, and the overhead for information distribution is also significant. In neighbor approach, each vehicle in a target area exchanges geocast messages and location information. Since this approach consists of only vehicles without infrastructure, many researchers have proposed various methods for practical use. As a method for determining the optimal target area range, [8] has proposed a method for analyzing periodically collected mobility information using a Software Defined Network (SDN), and while [9] has proposed an analysis method using a Convolutional Neural Network (CNN). In [10], an application for exchanging information between vehicles has been proposed assuming the infrastructure is unavailable during a disaster. [11] has proposed an efficient data delivery method by exchanging navigation information and predicting the vehicles moving toward the target area. In Floating Content [12] and Locus [13], vehicles exchange data lists with neighboring vehicles, send the data request if the data is not stored, and receive the data from neighboring vehicles. The vehicle with the data determines the transmission probability according to the distance from the point where the data is generated. The probability of data acquisition decreases as the vehicle moves further away from the point. On the other hand, when the vehicle density near the point is high, excessive data collision occurs because all vehicles transmit data with high probability. Furthermore, in these methods, for a user to obtain data, conventional query-response aware information distribution is required. Therefore, our STD-RS has introduced a mechanism that can efficiently deliver data to users while maintaining the data in the target area to promote the use of data around the point where data is generated.

3 STD-RS

In this section, we introduce the STD-RS [2–5], which is the basis of this paper.

3.1 Assumptions

The vehicles have wireless communication devices with conforming to the IEEE 802.11p communication standards, computing resources, and a global positioning system (GPS). Each vehicle also periodically transmits a beacon containing its identification (ID). In addition, the STD for the retention system has other information such as the center coordinate and radius of the retention area, the data transmission time, the data transmission interval, and the time-to-live (TTL).

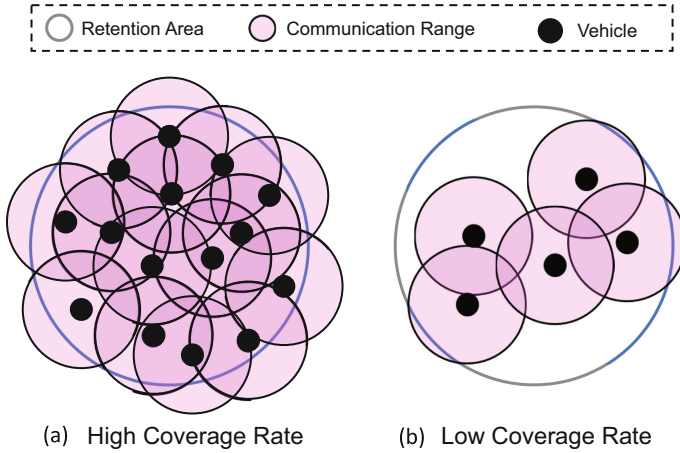


Fig. 1. The coverage rate.

3.2 System Requirements

In the conventional STD-RS, the vehicle broadcasts an STD and directly provides the STD to the user in the retention area. Therefore, it is necessary for data to reach the entire retention area by vehicular wireless communication. When the retention area is larger than the wireless communication range of vehicles, cooperation of multiple vehicles is essential. In addition, the vehicle must periodically broadcast STD so that the user can obtain the STD at any time within its TTL. Therefore, in the STD-RS, each vehicle must continue to broadcast STD into the retention area at the data transmission interval until its TTL expires, even for a moving user. We define the coverage rate in Eq. (1) as a metric of performance for the STD-RS:

$$CoverageRate = \frac{S_{EA}}{S_{RA}} \tag{1}$$

where S_{TA} is the size of the retention area and S_{EA} is the sum of the areas that the user can obtain data transmitted from any vehicle within the transmission interval.

Figure 1 shows an example of the coverage rate. The blue circle indicates the retention area, the black dot indicates a vehicle transmitting the data, and the pink circle indicates the communication range of a vehicle. As shown in Fig. 1(a), a high coverage rate means that users can passively obtain STD wherever they are in the retention area. On the other hand, as shown in Fig. 1(b), the low coverage rate means that users are not easy to obtain the STD within the retention area. Therefore, it is important to maintain a high coverage rate in STD-RS. If all vehicles transmit data at random in an environment where vehicle density is high, data collision occurs frequently, and the performance of data retention decreases. Therefore, the STD-RS must maintain a high coverage rate with minimum data transmission.

3.3 Previous Transmission Control Methods

To satisfy the above system requirements, we have proposed a method for controlling the data transmission probability using the number of neighboring vehicles and the number of received data [2,3]. In this method, each vehicle estimates the number of neighboring vehicles based on the received beacons from neighboring vehicles in the retention area. Each vehicle also measures the number of received data not beacons. The data transmission probability is set based on the number of neighboring vehicles and the number of received data for each data transmission interval. In the previous study [4], to prevent a decrease in the coverage rate in an environment where the vehicle density is low, the inter-vehicle distance was estimated from the received signal strength of data from the nearest vehicle and the data transmission interval was adjusted based on the distance. In this method, the data transmission probability is controlled according to the density of neighboring vehicles, thereby suppressing the increase in data transmissions while maintaining a high coverage rate. However, this method requires all vehicles to transmit beacons, resulting in the increased beacon transmissions, causing data collision, and consequently wasting wireless resource as the number of vehicles increases. In addition, since this method does not consider the attenuation, the reflect, and multi-path fading of radio waves due to obstacles, the fluctuation of radio waves due to Doppler effect caused by vehicle movement, and so on, a vehicle might not be able to transmit data even if radio waves from neighboring vehicles do not sufficiently cover its surroundings. Therefore, a new transmission control method that does not transmit beacons but considers the effects of the radio wave propagation environment is required to realize more effective data retention.

4 The Proposed Method

In this section, we propose a location-based autonomous transmission control method that is beacon-less and considers the radio wave propagation environment for the STD-RS.

4.1 Setting of Appropriate Transmission Positions

First, the proposed method sets an appropriate transmission position (ATP) that can provide STD over the entire retention area with the minimum number of data transmissions. Here, it is assumed that each vehicle installs common wireless communication interface with conforming to the IEEE 802.11p communication standards and know the communication range. The center of the retention area coincides with the center of a regular hexagon, which is approximated by the circumscribed circle as shown in Fig. 2. The radius in the circle represents the transmission range of the transceiver (e.g., IEEE 802.11p). Each vehicle can calculate ATP from hexagonal topology if the center coordinates of the STD and the radio communication distance are known. If vehicles can transmit data

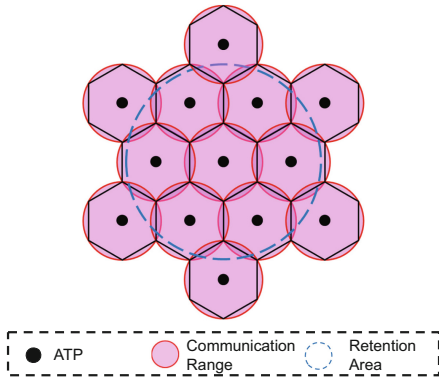


Fig. 2. ATP.

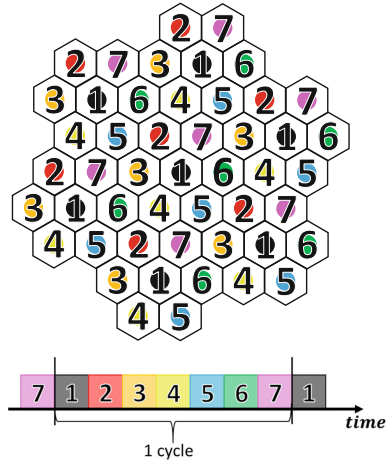


Fig. 3. Grouping for data transmission timing control.

on all ATPs in Fig. 2, they can cover the entire retention area with the minimum number of data transmissions. However, in the real traffic environment, vehicles does not always exist on the ATP because the vehicle moves. In addition, the actual communication range becomes shorter than ideal due to the attenuation of radio waves, etc. The proposed method controls the data transmission based on the received signal strength to address this problem in Sect. 4.4.

4.2 Grouping for Data Transmission Timing Control

Because vehicles in the neighboring hexagons tend to be hidden terminals, data collisions may occur. Therefore, the proposed method divides transmission points into groups to prevent data collisions. When data is transmitted based on a regular hexagon with an ATP set, it is divided into seven groups as shown in Fig. 3. In our proposed method, all vehicle with the wireless communication interface with conforming to the IEEE 802.11p share the same frequency band in order to communicate each other. In order to prevent the collision between different cells based on ATPs, the distance of two transmitting vehicles is greater than the transmission range. Therefore, the proposed method sets seven groups. If vehicles in different hexagons but of the same color group, the data transmissions are guaranteed not to collide as they are beyond the co-channel distance. Next, data is transmitted at different timings for each group. As shown at the bottom of Fig. 3, seven slots are created within the data transmission cycle, and the order of each slot is assigned to each group. The vehicle can calculate these slots from the data transmission time and transmission interval contained in the data (STD). The above method makes it possible to completely prevent data collisions with vehicles in the vicinity of other ATPs because data transmission

times differ for each group. However, since multiple vehicles are in the same hexagon, their conflict avoidance will be described in the next section.

4.3 Determination of Data Transmission Time in the Hexagon

As described in Sect. 4.1, the vehicle closer to the ATP can provide better coverage than other vehicles; so, it is better to give the data from the vehicle close to the ATP a high priority. The proposed method sets the data transmission time to each vehicle in a hexagon according to the following procedure. In this proposed method, all vehicles are supposed to be able to synchronize time using the GPS. First, each vehicle obtains its current location and from the GPS and calculates the distance D_p to the nearest ATP. Next, if the group of hexagons is the data transmission timing described in Sect. 4.2, the vehicle uses the following equation to set the data transmission time.

$$DataTransmissionTime = \frac{D_p}{D_v} * \frac{d}{7} + CurrentTime \quad (D_p \leq D_v) \quad (2)$$

where D_v is the radius of a regular hexagon, and d is the data transmission interval and $\frac{d}{7}$ is one slot time set in Sect. 4.2. The data transmission time is not set if D_p is larger than D_v . In other words, within each regular hexagon, the data generated by the vehicles is transmitted in the order of how close they are to the ATP.

4.4 Data Transmission Control Based on the Receiver Signal Strength Indicator (RSSI)

This section describes a data transmission control method that considers the attenuation of radio waves in each hexagon. Our proposed method uses the RSSI to control data transmission similar to [14, 15]. First, each vehicle in a hexagon broadcasts one STD at the data transmission time determined in Sect. 4.3. If the STD has not arrived from other vehicles by the transmission time in the same cycle, the vehicle broadcasts the STD. On the other hand, if a vehicle receives the STD in the same cycle, before its transmission time, it checks the RSSI, which is compared with the ideal RSSI estimated from the distance between the ATP and itself using the free space path loss model given in Eq. (3).

$$P_r = G_t G_r P_t \left(\frac{\lambda}{4\pi D_p} \right)^2 \quad (3)$$

where P_r is the Rx power, G_t is the Tx gain, G_r is the Rx gain, P_t is the Tx power, and λ is the wavelength. If the calculated RSSI exceeds the ideal RSSI, the vehicle determines that the STD has reached its surroundings and does not rebroadcast. On the other hand, if the calculated RSSI is smaller than an ideal RSSI, The vehicle determines that the STD has not reached the area to be covered its own, and rebroadcasts the STD. Our proposed method.

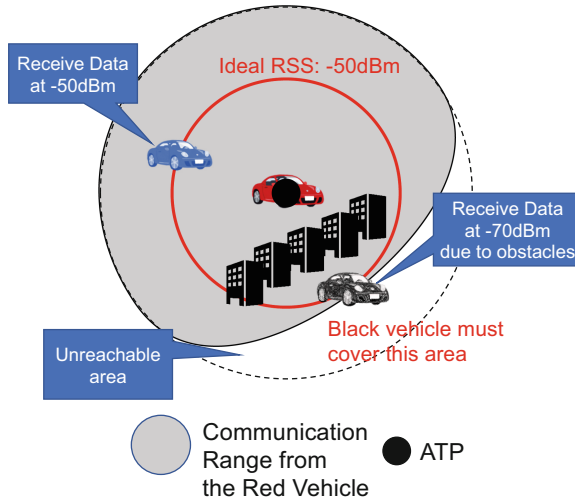


Fig. 4. Data transmission control based on RSSI.

The flow of the proposed method will be described with an example shown in Fig. 4. Here, it is assumed that the red vehicle broadcasts data at near the ATP. Furthermore, blue and black vehicles are the same distance from the ATP, and the ideal RSSI of the black and blue vehicles from the ATP is assumed to be -50 dBm. If the blue vehicle receives data at -50 dBm from the red vehicle, which is about the same as the expected value, the blue vehicle can judge that the radio wave of the red vehicle is ideally received. Therefore, the blue vehicle does not rebroadcast the data because the blue vehicle judges that there is no effect of the obstacles around its own. On the other hand, the black vehicle receives data at -60 dBm due to the obstacles. The black vehicle determines that the signal from the red vehicle does not reach beyond its own due to obstructions, etc. Thus, the black vehicle rebroadcasts data. Therefore, the black vehicle may be able to cover areas where radio waves have not reached.

5 Simulation Evaluation

In this section, we evaluate the effectiveness of the proposed method by simulation.

5.1 Simulation Model

We evaluated our proposed method using the Objective Modular Network Testbed in C++ (OMNeT++) network simulator [16], the Simulation of Urban MObility (SUMO) road traffic simulator [17], and the Vehicles in Network Simulation (Veins) framework, which implements the IEEE 802.11p specification for wireless communications [18]. Luxembourg SUMO Traffic (LuST) [19] was used

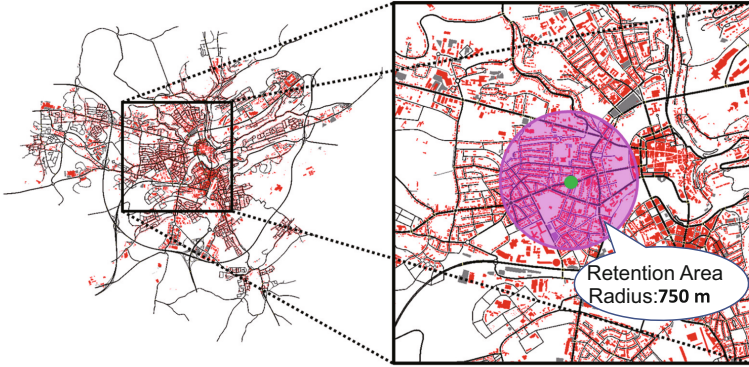


Fig. 5. The simulation area in LuST.

Table 1. The number of vehicles per simulation start time in LuST.

Simulation start time	The number of vehicles in the retention area
5:00 am	25
6:00 am	120
7:00 am	300
8:00 am	450

as a traffic model for vehicle mobility and signals. LuST will simulate the 24-h traffic flow of vehicles in Luxembourg. The simulation area includes expressways, arterial roads, and residential roads. The number of vehicles traveling on the road changes according to the time of day in the same way as the actual traffic. Each vehicle changes its route dynamically according to traffic conditions, and the operating time of traffic signals also varies according to traffic conditions. In this simulation, we set up a retention area in the center of Luxembourg, as shown in Fig. 5. The radius of the retention area is 750 m and the TTL is two minutes for the STD. The four times shown in Table 1 are used as the retention start time. The maximum transmission distance of the vehicle is 300 m, the data transmission interval is 5 s, the data size is 1000 bytes, and the transmission rate is 6 Mbps. In the simulation study, the proposed method is compared with naive method and our proposed method [2]. In the naive method, all vehicles re-broadcast the received STD, similar to the flooding. Thus, the naive method achieves the highest coverage rate among the methods in this simulation but has the most significant number of STD transmissions. The parameters in the data transmission probability control of the previous method are set as follows based on [2]; the beacon transmission interval is 5 s, the moving average coefficient is 0.5, and the target value β is 4.

5.2 Simulation Results

In this evaluation, an STD with a TTL of two minutes is retained. Five simulations were performed for each simulation start time of LuST in Table 1. The following results show the coverage rate, the number of STD transmissions, and the number of data collisions from one minute after the start of retention to the TTL. This is because of a transient state in which STD is diffused in the retention area for one minute from the beginning of retention. The following results show the state in which STD is steadily retained after diffusion.

Coverage Rate: Figure 6 shows the average coverage rate for each data transmission interval. The naive method's coverage rate is approximately 93% at 5:00 am when the traffic was light. The result shows that effective data retention is not achieved in an environment where only about 25 vehicles exist in the retention area. The proposed method's coverage rate is lower than the other two methods because the proposed method is different from the other two methods in that the STD does not reach most of the vehicles in the retention area in the diffusion. A diffusion method of STDs in our proposed method is future works. On the other hand, at 6:00 am, the coverage rate approach 99% for all methods, and after that, the all coverage rate is almost approximately 100%. These results indicate that at least 120 vehicles are required to effectively retain STD within a radius of 750 m in urban areas.

Total Number of STD Transmissions: Figure 7 shows the total number of STD transmissions. The naive method increase the number of STD transmissions with increasing traffic density. Both the previous and the proposed methods can suppress increasing of STD transmissions and converge to a nearly constant level even during the heavy traffic. During the heavy traffic (8:00 am), the number of STD transmissions of the previous method is approximately 1050 but the proposed method reduces it nearly to a half, approximately 500.

Total Number of Data Collisions: Figure 8 shows the total number of data collisions. Both naive and previous methods show an increase in the number of data collisions with an increase in traffic density. In particular, at 8:00 am, the previous method's data collisions are approximately 400 times higher than that of the proposed method. On the other hand, the proposed method causes almost no data collision regardless of the traffic density. The results showed that the proposed method could stably provide the STD to users.

From the above results, the proposed method can reduce the number of STD transmissions to a half of the previous method while achieving approximately 100% coverage rate even in heavy traffic without transmitting beacons. Moreover, since the proposed method can significantly reduce data collisions, it can realize an effective STD-RS even in a real traffic environment.

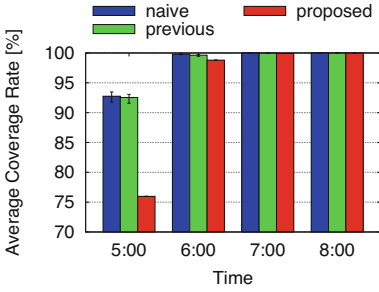


Fig. 6. Average coverage rate.

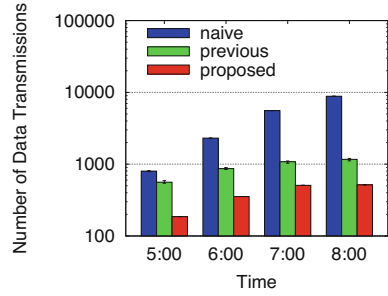


Fig. 7. The number of STD transmissions (semi-logarithm plot.)

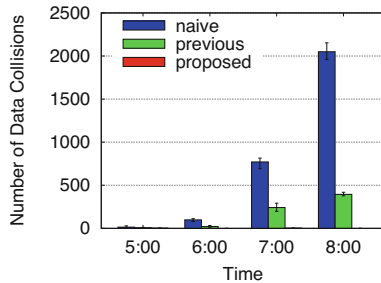


Fig. 8. The number of data collisions.

6 Conclusion

In this paper, to provide the STD to users effectively in a specific area while making efficient use of wireless resources, we propose a new transmission control method for the STD-RS. In the proposed method, the vehicle does not transmit the beacon. By controlling the data transmission based on the received signal strength of the data, it is possible to retain the STD effectively. Simulation results show that the proposed method can suppress the number of STD transmissions while achieving approximately 100% coverage rate in a time of heavy traffic in a natural environment. We believe that the proposed method can realize an effective STD-RS. In future works, we plan to study the retention system for multiple types of data and large-size data.

Acknowledgments. This work was supported in part by the Japan Society for Promotion of Science (JSPS) KAKENHI under Grants 20K11792, and the National Institute of Information and Communication Technology (NICT). Finally, we would like to express our appreciation to Ichiro Goto for his great contribution to this research.

References

1. Cisco Annual Internet Report (2018–2023). <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>
2. Nobayashi, D., Goto, I., Teshiba, H., Tsukamoto, K., Ikenaga, T., Gerla, M.: Adaptive data transmission control for spatio-temporal data retention over crowds of vehicles. *IEEE Trans. Mob. Comput.* Early Access (2021)
3. Teshiba, H., Nobayashi, D., Tsukamoto, K., Ikenaga, T.: Adaptive data transmission control for reliable and efficient spatio-temporal data retention by vehicles. In: Proceedings of ICN 2017, pp. 46–52, April 2017
4. Goto, I., Nobayashi, D., Tsukamoto, K., Ikenaga, T., Lee, M.J.: Transmission control method for data retention taking into account the low vehicle density environments. *IEICE Trans. Inf. Syst.* **E104-D**(4), 508–512 (2021)
5. Yamasaki, S., Nobayashi, D., Tsukamoto, K., Ikenaga, T., Lee, M.J.: Efficient data diffusion and elimination control method for spatio-temporal data retention system. *IEICE Trans. Commun.* **E104-B**(7), 805–816 (2021)
6. Maihofer, C.: A survey of geocast routing protocols. *IEEE Commun. Surv. Tutor.* **6**(2), 32–42 (2004)
7. Maihofer, C., Leinmuller, T., Schoch, E.: Abiding geocast: time-stable geocast for ad hoc networks. In: Proceedings of ACM VANET, pp. 20–29 (2005)
8. Maio, A., Soua, R., Palattella, M., Engel, T., Rizzo, G.: A centralized approach for setting floating content parameters in VANETs. In: 14th IEEE Annual Consumer Communications & CCNC 2017, pp. 712–715, January 2017
9. Manzo, G., Otalora, S., Braun, T., Marsan, M., Rizzo, G., Nguyen, H.: DeepFloat: resource-efficient dynamic management of vehicular floating content. In: 2019 31st International Teletraffic Congress (ITC 31), pp. 46–54 (2019)
10. Rizzo, G., Neukirchen, H.: Geo-based content sharing for disaster relief applications. In: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. *Advance in Intelligent System and Computing*, vol. 612, pp. 894–903 (2017)
11. Leontiadis, I., Costa, P., Mascolo, C.: Persistent content-based information dissemination in hybrid vehicular networks. In: Proceedings of IEEE PerCom, pp. 1–10 (2009)
12. Ott, J., Hyyti, E., Lassila, P., Vaegs, T., Kangasharju, J.: Floating content: information sharing in urban areas. In: Proceedings of IEEE PerCom, pp. 136–146 (2011)
13. Thompson, N., Crepaldi, R., Kravets, R.: Locus: a location-based data overlay for disruption-tolerant networks. In: Proceedings of ACM CHANTS, pp. 47–54 (2010)
14. Zhu, C., Lee, M.J., Saadawi, T.: A smart broadcast scheme for wireless military networks. In: Proceedings of IEEE Military Communications Conference (MILCOM 2004), pp. 251–257 (2004)
15. Zhu, C., Lee, M.J., Saadawi, T.: A border-aware broadcast scheme for wireless ad hoc network. In: Proceedings of IEEE Consumer Communications and Networking Conference (CCNC 2004), pp. 134–139 (2004)
16. OMNeT++. <https://omnetpp.org/>
17. SUMO. http://www.dlr.de/ts/en/desktopdefault.aspx/tabid-9883/16931_read-41000/
18. Veins. <http://veins.car2x.org/>
19. Codeca, L., Frank, R., Engel, T.: Luxembourg SUMO Traffic (LuST) scenario: 24 hours of mobility for vehicular networking research. In: 2015 IEEE Vehicular Networking Conference (VNC), pp. 1–8 (2015)



Vehicle Routing in Whole and Segmented Areas to Incrementally Collect the Disaster Information

Sanjukta Khwairakpam^(✉), Masahiro Shibata, and Masato Tsuru

Computer Science and Systems Engineering, Kyushu Institute of Technology,
Fukuoka, Japan

sanjukta.khwairakpam429@mail.kyutech.jp,
{shibata,tsuru}@csn.kyutech.ac.jp

Abstract. In large-scale disaster situations, the emergency disaster response headquarters (HQ) is essential to incrementally collect the damage information from the whole area for early decisions and responses. We consider the scenarios in which such information should be monitored and brought to HQ by patrolling vehicles with sensing devices especially when the high-speed communications infrastructures are unavailable. Our previous work studied a routing problem for two vehicles that can return to HQ multiple times on the way to minimize the average delay time of incrementally collecting information. However, such a joint optimization of the vehicles' routes will be complicated in a large area. In this paper, therefore, we consider an area-segmentation approach in which each vehicle's route is designed for collecting one of sub-areas' information only. We show the impact of the area-segmentation on the average delay time for the information collection from the whole area.

1 Introduction

Large-scale disasters related to typhoon, earthquakes, rainstorms, hurricane, etc. occur every year and the frequency has increased over decades, killing millions of people and damaging their properties. When such a large-scale disaster happens, the disaster management team needs to establish an on-site emergency disaster response headquarters (HQ) for disaster management. For early and appropriate actions to be taken in the initial response and partial recovery, the HQ should promptly and incrementally collect the information from the overall region which may be damaged.

For disaster information collection, it is often considered to use patrolling vehicles that equip mics, cameras, and other sensors and cruise all streets in a region to monitor the disaster damage situation around each street. If the high-speed telecommunication network infrastructures are available, such information can be sent to HQ in an online manner. However, during or aftermath of disaster, such telecommunication infrastructures may be unavailable or degraded,

Supported by NICT, Japan (No. 22007) and JSPS KAKENHI Grant Number21K17706.

and thus, each patrolling vehicle should move not only to monitor the damage information along streets but also to bring that information to HQ by itself.

Therefore, we consider a kind of Arc Routing Problem in Vehicle Routing Problem (VRP) [1, 2], in which one or more vehicles start from HQ, monitor the information along all the streets and bring back the information to HQ. Chinese Postman Problem (CPP) [3] is a well-known Arc Routing Problem and applied to real-world applications such as optimal routing for garbage collection [4]. Although we borrow the idea of the CPP, our setting is clearly different. While the CPP and conventional Arc Routing Problems mainly aim to minimize the time and the cost taken to finally deliver or collect all the necessary items, we focus on minimizing the average delay time of incrementally collecting the overall information by allowing vehicles to return to HQ multiple times on the way to drop a partial monitored information.

In our initial studies [5, 6], we systematically searched for good routes of one or two vehicles to minimize the average delay time of incrementally collecting the whole area information. However, since it requires a joint and global optimization in finding good travel routes, its computational cost will be large as the whole area becomes large with a number of vehicles. Therefore, this present paper investigates an area-segmentation approach in which the whole area is segmented into multiple sub-areas, each vehicle is designated to one of sub-areas, and its route is designed for collecting the sub-area's information only. The impact of the area-segmentation for information collection is discussed through some examples. We consider the situation not only when all streets are normal but also when some streets are damaged but it is not known by traveling vehicles.

2 Model

To model a street map in a region, an undirected connected network graph is used. In this paper, the terms “link” and “node” are used instead of “edge” and “vertex”. On an undirected graph, the degree of a node is defined as the number of links the node has. The emergency disaster response headquarters (HQ) is located at a node of the network and is identically treated as that node. On the network, one or more patrolling vehicles start from HQ at the same time and return to HQ to bring monitored information; each link should be passed at least once by some vehicle; and each vehicle can return to HQ multiple times on the way to drop the monitored information. A route (or a travel route) of a vehicle is the sequence of links on which the vehicle travels from HQ and finally returns to HQ. The term “length” means the number of links in a sequence of links such as a path, a circuit, and a route. If the degree of HQ is k , i.e., HQ has k links, the vehicles need to return to HQ at least $\lceil k/2 \rceil$ times in total to cover those k links on their routes. Let R be the total number of times that the vehicles return to HQ. We focus on finding a good set of travel routes of vehicles based on the following two criteria:

1. Last Information Delay-time (LID): the time when the entire information of all links has been brought to HQ by the vehicles, i.e., the last returning time of the vehicles for information collection.

2. Information Delay-time Product (IDP): defined as

$$\int_0^{LID} (1 - u(t)) dt$$

where $u(t)$ is the ratio of the amount of the information brought to HQ by the vehicles until time t to the amount of the entire information; t is the time taken from the start. The IDP represents the average delay time for information collection, i.e., the product of the fraction of the information and its delay time in delivering to HQ. A smaller IDP benefits early decisions and responses by HQ.

To make “LID” small, the travel route from HQ to HQ should be as short as possible. On the other hand, to make “IDP” small, the shortest route is not always a good choice as suggested from the experimental results in Sect. 6.

The following two assumptions are used to make our model simple. Note that a node does not necessarily mean an intersection of streets. To be consistent with the following assumptions on links, a number of nodes can be arranged on a “long” or “information-dense” street.

1. The unit time is taken as a vehicle passes through a link in one direction between two nodes. In other words, the number of links passed by a vehicle is equal to the time taken for the vehicle’s travel.
2. The unit amount of information corresponds to one link. Since the information associated to a link is unchanged, if a link is passed twice by the same or different vehicles, the same (duplicated) information is monitored. Hence, $u(t)$ is the ratio of the number of different links passed by the vehicles returning to HQ before time t to the total number of links of the entire network.

3 Simple Examples

We consider a street network consisting of 12 links and 9 nodes shown in Fig. 1(a). A graph that includes even degree nodes only is called an Euler graph. If a graph is an Euler graph, from any starting point, a single vehicle can traverse all links and return back to the starting point very efficiently, i.e., it can pass through every link once and only once. Such a travel route is called an Eulerian circuit. On the other hand, if a graph is not an Euler graph as shown in Fig. 1(a), by virtually adding the duplicate links between the pairs of odd degree nodes, the given graph can be converted to an Euler graph as shown in Fig. 1(b) and (c), while the total number of links on the graph is increased. Any Eulerian circuit on the converted graph is also an efficient travel route of a vehicle if the number of additional virtual links is small. This is the basic idea behind the Chinese Postman Problem mentioned in Sect. 1. Note that, any Eulerian circuit in Fig. 1(b) and (c) should pass the center node of this graph twice ($R = 2$) and four times ($R = 4$), respectively.

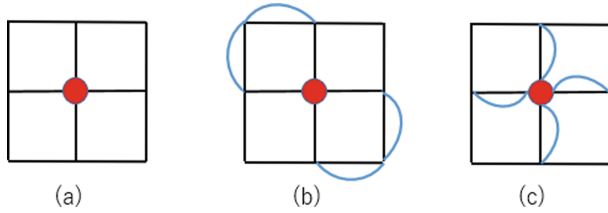


Fig. 1. (a) An undirected connected graph, (b) an example of converted Euler graph, (c) another converted Euler graph.

In the following examples, HQ is located at the center node and two vehicles start from HQ at the same time and return to HQ to bring the collected information along the links. We examine different good routes by changing R , the total number of times that the vehicles return to HQ. A good pair of routes of two vehicles for $R = 2$ is shown in Fig. 2(a). Vehicles A and B travel along the circuits indicated by red and green lines, respectively. They bring back the monitored information to HQ at time 8 simultaneously. Each vehicle monitors a half of the entire graph, HQ gets all information at time 8 (i.e., $LID = 8$).

For $R = 4$, two cases are shown in Fig. 2(b) and (c). In both cases, vehicles A and B travel along the two-round circuits indicated by red and green lines, respectively; a solid line is for the first round and a dot line is for the second round of a route. They return to HQ first at time 4 simultaneously to drop partial information monitored before, and return to HQ last at time 8 to bring the remaining information so that the LID is 8. However, the fraction of the information monitored by two vehicles at time 4 is different in two cases; $\frac{7}{12}$ in (b) and $\frac{8}{12}$ in (c), respectively. This is because the first round circuit for vehicle A and that for vehicle B are overlapped in case of (b) but not overlapped in case of (c). Note that, in case of (b), each vehicle can be seen as traveling on the whole area or traveling on two of one-quarter sub-areas separately. On the other hand, in case of (c), each vehicle can be seen as traveling on each of half sub-areas separately. This is the view point of the area-segmentation explained in Sect. 5 later.

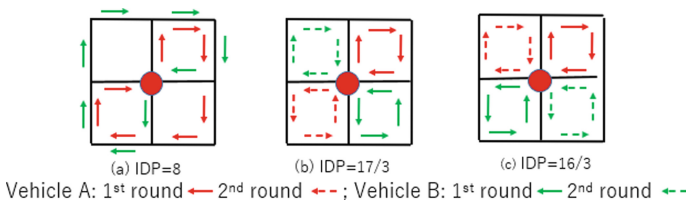


Fig. 2. Good routes of two vehicles for (a) $R = 2$ (left), (b) a case of $R = 4$ (center), and (c) another case of $R = 4$ (right).

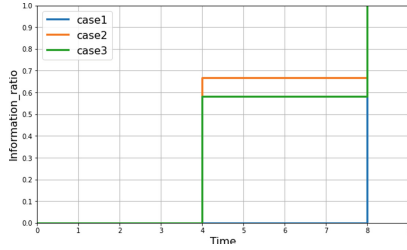


Fig. 3. Time evolution of $u(t)$: the fraction of the information brought to HQ until time t

Figure 3 shows $u(t)$ defined in Sect. 2 (the fraction of the information brought to HQ until time t) of the above exemplified routes in Fig. 2(a), (b), and (c) indicated by blue, orange, and green lines, respectively. From Fig. 3, the IDPs of the routes in cases of (a), (b), and (c) can be computed as 8, $17/3 = 5.67$, and $16/3 = 5.33$, respectively. Note that $IDP = 8$ is the smallest in condition of $R = 2$ and $IDP = 16/3$ is also the smallest in condition of $R = 4$. The examples suggest that multiple returns to HQ to drop the monitored information can reduce the IDP.

4 Searching for Good Routes on an Area

For a given undirected connected graph and a node as HQ on that graph, we search for good routes of one or two vehicles starting from and ending at HQ to cover all links in total. We propose a systematic search consisting of the following steps.

1. Set the number R :
Choose any number R of return times to HQ as long as R is equal to $\lceil k/2 \rceil$ or more where k is the degree of HQ.
2. Add virtual links to HQ depending on R :
If $2R > k$, create an extended graph by adding $(2R - k)$ virtual links to HQ randomly each of which connects HQ and one of its k neighboring nodes on the network graph.
3. Make an Euler graph:
List all the odd degree nodes in the extended network graph. Note that the number of the odd degree nodes is always even. Let $2m$ be that number.
If $m \geq 1$, make m pairs from $2m$ odd degree nodes and select m paths that connect paired odd degree nodes on the graph randomly so that the total length of connecting paths (the total number of links along those paths) is as small as possible. Note that HQ should not be passed across by any connecting path so as that the value R is unchanged. Then add the virtual duplicate links along those connecting paths to make the entire graph an Euler graph (i.e., to make the degree of each node even).

4. Find an Eulerian circuit:

On the Euler graph, find an entire Eulerian circuit randomly. The entire Eulerian circuit of the graph is composed of sub-circuits C_1, C_2, \dots, C_R including HQ where each sub-circuit passes HQ once and only once. Without loss of generality, we assume $|C_i| \leq |C_j|$ for $(i < j)$ where $|C_i|$ is the length of sub-circuit C_i .

5. Get the travel route(s) of vehicle(s):

For a single vehicle, the sub-circuits C_1, C_2, \dots, C_R are concatenated (in any order) to get a single route starting from HQ and returning to HQ R times. For two vehicles, each sub-circuit is assigned to one of two vehicles.

– If $R = 2$, vehicle A travels along C_1 , and vehicle B travels along C_2 .

– If $R \geq 3$, the sub-circuits C_1, \dots, C_R are divided into two groups G_A and G_B . Vehicle A travels along a route of concatenating all sub-circuit(s) in G_A (in any order), and Vehicle B travels along a route of concatenating all sub-circuits in G_B . In this grouping, the balance of the total length of sub-circuits in G_A and that in G_B is taken into account so that the difference between the lengths of two routes is not large.

Let P^* be the obtained travel route(s) of vehicle(s) on the given Eulerian circuit.

6. Compute LID and IDP of travel routes P^* :

Compute LID (Last Information Delay time) and IDP (Information Delay-time Product) of P^* by taking account of the length of each sub-circuit in $\{C_1, C_2, \dots, C_R\}$ and their overlapping. Note that IDP of P^* is not for the converted Euler graph but for the original graph.

7. Select the best P^* on the given Eulerian circuit:

There are different possible P^* s in Step 5 depending on the grouping and the concatenation of the sub-circuits. Among all P^* s that minimize IDP, select one that has a smallest LID. This is the best P^* for an Eulerian circuit in Step 4.

8. Find candidates of good P^* across different choices:

Repeat Steps 1 to 7 by changing the choices of the number R , the virtual duplicate links (i.e., the Euler graph), and the Eulerian circuit, we finally get an appropriate candidate set of travel routes in terms of IDP and LID.

5 Searching for Good Routes on Segmented Areas

The procedure explained in Sect. 4 systematically searches for good routes of patrolling vehicles to minimize the IDP for information collection from a given area. However, applying that procedure to the whole area (called “the whole-area approach”) is computationally costly when the whole area is large and the number of vehicles is not small. This is because the procedure leverages a variety of Euler graphs and Eulerian circuits on them and requires a joint combinatorial optimization in finding a good set of travel routes of vehicles. On the other hand, as shown in simple examples in Sect. 3, on some of good routes (i.e., small IDP routes), each vehicle can be seen as traveling on each of segmented areas separately. Therefore, we consider “the area-segmentation” in which the whole area is segmented into multiple sub-areas, each vehicle is designated to one of

sub-areas, and its route is designed for collecting the sub-area’s information only. The area-segmentation approach may be able to only find semi-optimal routes but is expected to be able to much reduce the computational cost. More exactly, the area-segmentation approach consists of the following steps; the number of vehicles is assumed to be 2 here.

1. Segment the whole area:
The whole area is segmented into two sub-areas with almost equivalent numbers of links so that the lengths of two vehicles’ routes are not much different.
2. Search for good single vehicle’s routes for each sub-area:
Using the procedure proposed in Sect. 4, a candidate set of good routes for a single vehicle is obtained in each segmented sub-area by regarding a sub-area as the whole area. Any obtained route for a sub-area is a good route of single vehicle in terms of IDP and LID on its sub-area only.
3. Select good pairs of sub-area routes:
For the top m good routes for each sub-area obtained in Step 2, compute IDP and LID on the whole area for each pair in all the $m \times m$ combinations of two routes on two sub-areas, and select the best one or a few of good ones.
4. Find candidates of good area-segmented routes across different segmentations:
Repeat Steps 1 to 3 by changing the area segmentation, we finally get an appropriate candidate set of routes in terms of IDP and LID on the whole area.

6 Experimental Results

6.1 Normal Grid Streets

A 5×5 grid map is considered as a town street network in the normal condition (without any damaged streets) consisting of 40 links and 25 nodes shown in Fig. 4(a). HQ is located at a center node with degree of 4. We consider two

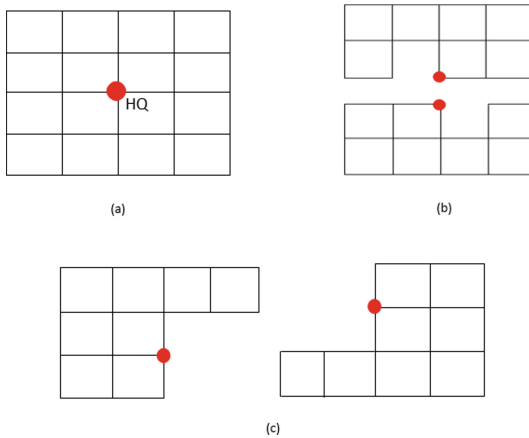


Fig. 4. (a) The whole area, (b) the segmented areas of Pattern 1 and (c) Pattern 2.

vehicles starting from HQ and ending at HQ with $R = 4$, i.e., two vehicles return to HQ four times in total. To investigate the impact of the area segmentation on the information collection performance, we compare the good routes in terms of IDP and LID in the whole area approach and two examples of the segmentation approach with different segmentation patterns (Patterns 1 and 2) shown in Fig. 4(b) and (c). The procedure in Sect. 4 and 5 was implemented in Python and was conducted 100000 times for the whole area and each sub-area.

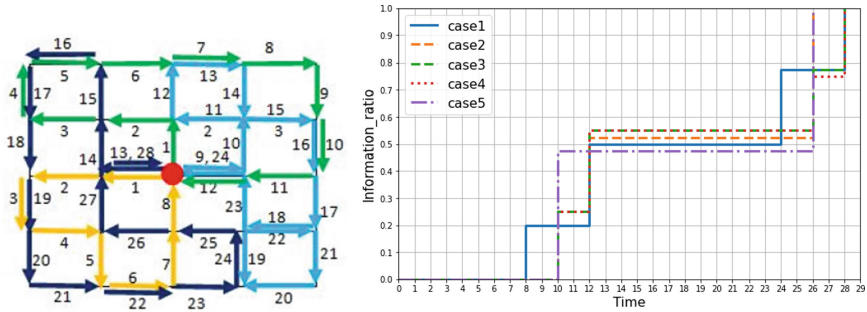


Fig. 5. One of the best routes (left) and the good five cases of $u(t)$ (right) in the whole area approach ($R = 4$, two vehicles).

For the whole area approach, Fig. 5(left) shows one of the best routes (Case 1); vehicles A and B start at the same time and return to HQ at time 8 and 12 to drop a partial information for the first round trip indicating yellow and blue lines, again start travel and return to HQ at time 24 and 28 to drop all remaining information. In Fig. 5(right), $u(t)$ of the five good cases are shown. Case 1 has the smallest IDP ($= 18.1$) while Case 2 and Case 5 have the same smallest LID ($= 26$).

For the segmentation approach Pattern 1, Fig. 6 shows one of the best routes (Case 1) and $u(t)$ of the five good cases. Suppose the vehicle A collects information in segment 1 and returns to HQ twice (the upper half) and vehicle B does the same in segment 2 (the lower half). As explained in Sect. 5, to find good pairs of routes of two vehicles without a joint and global optimization, we search for ten good routes of vehicle A only in segment 1 and those of vehicle B only in segment 2 independently. Then, by combining the routes of two segments together, we select five good pairs of routes with first five smallest IDPs. Case 1, indicating by blue line in Fig. 6(right), has the smallest $IDP(= 19.2)$ and $LID = 30$. In Case 1, shown in Fig. 6(left), two vehicles return to HQ at the same time of 12 with a partial information (the fraction is $24/40 = 0.6$) and again return to HQ with the remaining information at time 30.

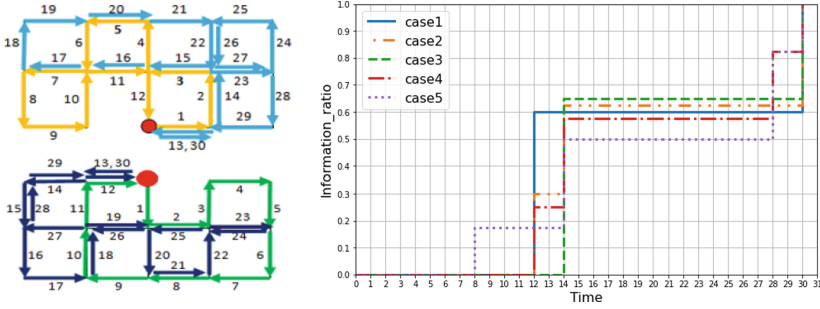


Fig. 6. One of the best routes (left) and the good five cases of $u(t)$ (right) of the segmented approach Pattern 1 ($R = 4$, two vehicles).

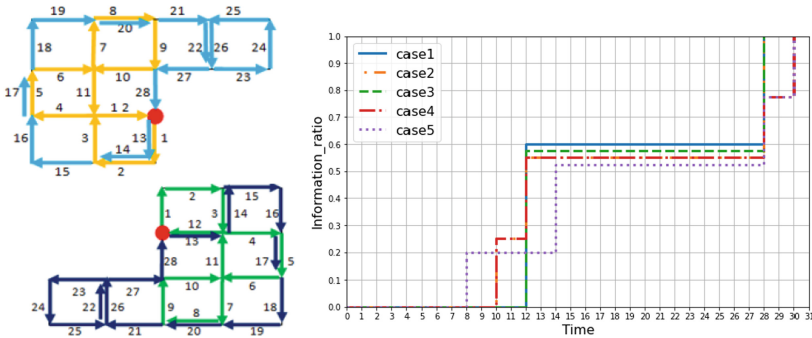


Fig. 7. One of the best routes (left) and the good five cases of $u(t)$ right of the segmented approach Pattern 2 ($R = 4$, two vehicles).

For the segmentation approach Pattern 2, Fig. 7 shows one of the best routes (Case 1) and $u(t)$ of the five good cases. In the same way as Pattern 1, we get ten good routes with small IDPs in each segment, and by combining together, we select the five good pairs of routes with first five smallest IDPs. There are two different LID values, 28 and 30. Case 1, indicating by blue line in Fig. 7(right), has the smallest $IDP(= 18.4)$ and the smallest $LID = 28$. In Fig. 7(left), two vehicles return to HQ at the same time of 12 with a partial information (the fraction is $24/40 = 0.6$) and again return to HQ with the remaining information at time 28.

6.2 Partially Damaged Grid Streets

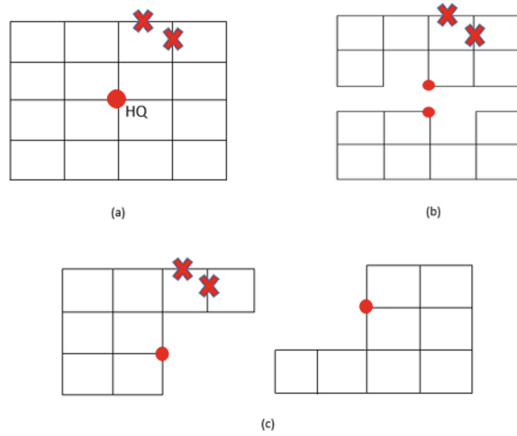


Fig. 8. The whole area (a) and the segmented areas (b) and (c) with two damaged links.

When the disaster happens, some streets may be damaged through which vehicles cannot pass quickly; it will take a long time to pass through such a street. To investigate the impact of the area segmentation with damaged streets, we check the information collection performance degradation of five good cases when some damaged streets (links) require a much longer time to pass through but those streets are not known by vehicles traveling the given good routes. Two damaged links are placed as shown in Fig. 8(a), (b), and (c); such damaged links are likely to happen side-by-side due to the proximity of damages in a disaster situation. We assume it takes 4 unit times when a vehicle passes each damaged link.

For the five good cases in each of the whole area approach and the segmented Patterns 1 and 2 in the normal condition, we examine how their IDPs and LIDs are degraded (increased) in the damaged condition as shown in Table 1. In each, its IDP and LID must be increased in the damaged condition but the difference strongly depends on the case. In the whole area approach, the best case (Case 1)'s IDP is increased from 18.1 to 21.32, while Case 4's IDP is increased from 18.3 to 20.25, i.e., Case 4 has the best IDP in this specific damaged condition. Similarly, Cases 2 and 5 have the smallest LID in the normal condition but Case 1 has the smallest LID in the damaged condition. In general, when a vehicle meets a damaged link and slowly passes the link on its route, its adverse impact on IDP becomes larger as the meeting of the damaged link happens at an earlier time, while LID is simply increased by 3 if that vehicle comes back to HQ last in the normal condition.

In the segmentation approach, it is also seen that the best case in the normal condition is not necessarily the best in the damaged condition. Furthermore, in

Table 1. The impact of damaged links on the IDPs and LIDs of the five good cases in each approach.

Case	(a) Whole area approach				(b) Segmented (Pattern 1)				(c) Segmented (Pattern 2)			
	Normal		Damaged		Normal		Damaged		Normal		Damaged	
	IDP	LID	IDP	LID	IDP	LID	IDP	LID	IDP	LID	IDP	LID
1	18.1	28	21.32	31	19.2	30	20.32	39	18.4	28	19.75	37
2	18.15	26	22.27	35	19.4	30	20.07	39	18.7	28	20.5	37
3	18.25	28	21.02	35	19.6	30	22	39	18.8	28	20.82	37
4	18.3	28	20.25	32	19.8	30	22.8	34	19.15	30	22.13	35
5	18.4	26	20.65	35	20.3	30	22.18	34	19.9	30	21.98	37

terms of IDP, the advantage of the whole area approach to the segmentation approach seems unclear in the damaged condition. The smallest IDPs in the damaged condition are 20.25 (Case 4), 20.07 (Case 2), and 19.75 (Case 1) in the whole area approach, Pattern 1, and Pattern 2, respectively; the differences are small and the order is reversed compared with the normal condition.

6.3 Discussion

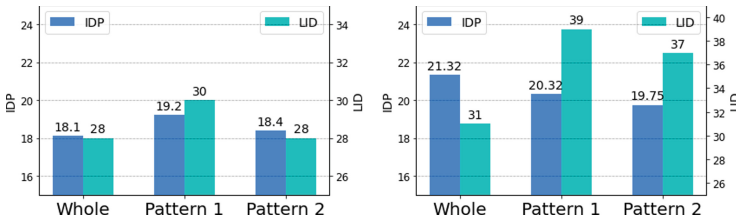


Fig. 9. IDP/LID of good cases by three approaches in the normal condition (left) and their degraded values in the damaged condition (right)

We focus on the best IDP case in each of the whole area approach and the segmented Patterns 1 and 2 in the normal condition. Figure 9 compares IDPs and LIDs of those three cases in the normal condition (left) and those values in the damaged condition (right). It can be seen that the best case in the whole area approach exhibits a larger increase in IDP with a smaller increase in LID, while the best cases in the segmented Patterns 1 and 2 exhibit a smaller increase in IDP with a larger increase in LID. Note that, in our configuration, two damaged links are placed only within one sub-area, and thus only one vehicle is affected by all two damaged links in Patterns 1 and 2. Therefore, for a pair of routes in the segmentation approach, it can be expected that the damaged links have a less impact on increasing the average delay time in collecting the whole information

and a more impact on increasing the last vehicle's returning time. This discussion suggests that a good pair of routes obtained by the segmentation approach is robust in terms of IDP compared with the whole area approach if using two or more vehicles for information collection.

7 Conclusion

We have presented a kind of Arc Routing Problem in which one or more patrolling vehicles start from HQ, monitor the information along all streets and bring back the information to HQ. We focus on IDP that represents the average delay time for information collection, and search for good travel routes to minimize the IDP. In particular, we compare the whole area approach and the area-segmentation approach to search a good pair of routes of two patrolling vehicles. Although the area-segmentation approach restricts the search space and thus can only find semi-optimal routes, it is suggested that the segmentation will contribute to the robustness by mitigating the increase of IDP due to the damaged links.

In the present model, two vehicles collect the information independently, that is an independent collection scheme. From a methodological aspect, we are developing a collaborative collection scheme in which two vehicles can interact on the way to delegate the information collected by one vehicle to the other, which is expected to reduce the IDP while keeping the LID.

References

1. Toth, P., Vigo, D. (ed.): The Vehicle Routing Problem. SIAM Discrete Mathematics and Applications (2002)
2. Hsueh, C.-F., Chen, H.-K., Chou, H.-W.: Dynamic vehicle routing for relief logistics in natural disasters. *Veh. Routing Probl.* **1**, 71–84 (2008)
3. Grotschel, M., Yuan, Y.: Euler-Mei-Ko Kwan, Konigsberg, and a Chinese postman. *Documenta Mathematica: Optim. Stories* 43–50 (2012)
4. Buhrkal, K., Larsen, A., Ropke, S.: The waste collection vehicle routing problem with time windows in a city logistics context. *Procedia Soc. Behav. Sci.* **39**, 241–254 (2012)
5. Maki, Y., Mu, W., Shibata, M., Tsuru, M.: Vehicle routing for incremental collection of disaster information along streets. In: Hara, T., Yamaguchi, H. (eds.) *MobiQuitous 2021*. LNICST, vol. 419, pp. 487–492. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-94822-1_28
6. Khwairakpam, S., Shibata, M., Tsuru, M.: Vehicle routing to minimize the average delay for collecting the disaster damage information. In: *Proceedings of the IEEE WF-IoT 2022* (2022, to appear)



Towards a Methodology for the Semantic Representation of Iot Sensors and BPMNs to Discover Business Process Patterns: A Smart Irrigation Case Study

Beniamino Di Martino^{1,2,3}, Luigi Colucci Cante¹, Antonio Esposito^{1(✉)},
and Mariangela Graziano¹

¹ Department of Engineering, University of Campania L. Vanvitelli, Aversa, Italy
{beniamino.dimartino, luigi.coluccicante, antonio.esposito,
mariangela.graziano}@unicampania.it

² Department of Computer Science and Information Engineering, Asia University,
Taichung, Taiwan

³ Department of Computer Science, University of Vienna, Vienna, Austria

Abstract. In recent years, information technology has played a decisive role in farm management through the exploitation of smart sensors and IoT devices. The introduction of IoT has improved the entire agricultural process chain, from Smart Irrigation to Smart Seeding. Another interesting aspect regards the application of semantic and artificial intelligence techniques to these sectors. This work moves in this direction, providing a methodology for the implementation of an expert system helping the smart management of irrigation systems using an approach based on ontologies, BPMN semantic annotation and logical inference techniques. Through the Irrig ontology, proposed by the INRAE research centre as the knowledge base, the expert system aims at providing decision support for the automatic activation of actuators of smart irrigation systems, and verifying the compliance of farm business processes with the related regulations, using an approach based on the Business Process Patterns discovery in semantically annotated BPMNs.

1 Introduction and Motivation

Technological innovation is involving many sectors, leading to a real revolution in methodologies and tools, and consequently to significant economic growth. Even agriculture has been experiencing a technological revolution in recent years. For several years now, we have been listening to talks about **Smart Agriculture**, the aim of which is to combine traditional agriculture with new digital and technical solutions in order to facilitate the work of farmers and increase efficiency in the various daily activities of farms. With its numerous specialised software, applications, networks and sensors, the advent of technologies based on the **Internet of Things (IoT)** is capable of bringing new levels of productivity and efficiency even to a sector that has always been tied to traditional business

models. There are many problems that affect the agricultural sector, with the unpredictability of weather conditions perhaps the biggest problem to be faced, but until recent years, prevention and care aiming at good harvests were based only on observation and experience, which have always characterised farmers, but those are now not enough on their own. The help of IoT and Cloud-based technologies can form the basis for decision support systems that can guide farmers or modern equipment to intervene in the field only when it is really necessary, enabling the staff of **agri-businesses** to optimise their business processes. This paper focuses on one of the most important activities of Smart Agriculture: Smart Irrigation. In recent years, **Smart Irrigation Technology (SIT)** has been gaining ground, which, thanks to the use of intelligent and interconnected sensors in an IoT system, offers the possibility of improving the use of water resources in the prediction of climate change. SITs are remote control systems that allow on-demand irrigation to be set up, e.g. by activating it when humidity falls below a critical threshold and stopping it when the water status of the soil/plant system has returned to the optimum. This paper proposes a new approach based on the use of a semantic-based methodology for the implementation of decision support systems to manage Smart Irrigation Technology. The methodology involves the use of ontologies, semantic annotation of business processes represented using BPMN (Business Process Model and Notation) notation, and the construction of logic-based inference rules. Semantics is applied to provide two features: 1) automatic activation of irrigation systems when certain weather conditions are verified; 2) compliance verification of business processes used by farms for the intelligent management of irrigation activities, through the detection in the processes of appropriate business process patterns that sum up the relevant regulations. The paper is structured as follows: an introductory Sect. 1 on the paper that provides the motivation for it, a Sect. 2 on related work on studies on the application of IoT in the smart agriculture sector and on some work proposing an application of a semantic-based approach, followed by Sects. 3 and 4 on the proposed methodology and its implementation and realization of the expert system, and finally a Sect. 5 that concludes the paper with considerations on future work.

2 Related Works

In recent years, several works have been conducted to investigate the adoption and introduction of IoT devices in the field of agriculture, and some of them also proposed approaches involving semantics and artificial intelligence. IoT solutions are focused on helping farmers close the supply-demand gap, by ensuring high yields, profitability, and protection of the environment. The approach of using IoT technology to ensure optimum application of resources to achieve high crop yields and reduce operational costs is called precision agriculture. Some of the emerging technologies for intelligent agriculture based on the Internet of Things (IoT) are presented in [7]. The fast emergence of IoT-based technologies has redesigned the sector of smart agriculture which has shifted the industry from

statistical to quantitative approaches. These revolutionary changes are reshaping existing farming methods and creating new opportunities with their set of potential challenges. The article [1] focuses on the potential of wireless sensors and IoT in agriculture, as well as the challenges expected when integrating this technology with traditional farming practices. An example of the application of semantics and machine learning technologies to the field of smart agriculture is presented in [12], in which the EcoLoop project is described, which proposes an ICT system capable of collecting, aggregating and analysing IoT data, with the aim of promoting the reuse of wastewater and optimising water use in agriculture to propose a solution to one of the main problems facing agriculture. In this work, an interesting decision support system (DSS) is proposed that acts on wastewater systems by managing irrigation and fertilisation strategies, reservation queues and network distribution by exploiting smart sensors, semantic ontologies and machine learning technologies. Agriculture is an area where IoT applications have much potential. The market is full of devices that collect data from farms and send it to the cloud. The Semantic Web offers semantic interoperability that facilitates communication between heterogeneous devices and technology platforms. The work [8] discusses on the needs and requirements of IoT with semantic interoperability in agriculture. The use of IoT interoperability in agriculture can bring long-term benefits to farmers and increase productivity by reducing the overall costs incurred.

3 Description of the Methodology

The paper proposes a methodology for the development of an expert system for decision support and the search for business process patterns describing the smart management of irrigation in agricultural enterprises using a semantic approach. The aim of this expert system is to evaluate the status of an irrigation system with several sensors, determining the appropriate operation and compliance of the system with respect to a particular standard. The method applied for decision support is based on the **Irrinov method** [2], developed by Arvalis, which defines the best time to start irrigation, considering fundamental crop and soil properties. The methodology proposed is shown in Fig. 1. The first step of the methodology involves the definition of a business process, represented in BPMN, which describes all the business activities that a farm performs to intelligently manage the farms irrigation systems. Using the methodology described in the works [4, 5, 11], the BPMN is semantically annotated using domain ontologies that describe all concepts inherent to the farm, field irrigation, and IoT sensing. These ontologies are populated, using simple automated programs, with the values of the readings of the various sensors. The work described in [3] also provides a semantic annotation tool which makes it possible to define semantic relationships between a BPMN element and a concept in the ontologies. The output of the semantic annotation tool is an ontology called the BPMN-MM Ontology, which imports both a semantic representation of the structural aspects of the BPMN and the domain ontologies used, and contains all the semantic relations between the structural elements of the BPMN and the domain concepts

defined using the tool. This new ontology constitutes the Knowledge Base used for the construction of an expert system, which uses a logic rule-based inferential approach, which is why the BPMN-MM Ontology must be translated into a set of facts suitable for an appropriate inferential language. Then the user can prepare some inferential rules to be passed to an inferential engine, which can deduce some conclusions and infer new knowledge by working on the facts. The expert system has two main objectives: i) Provide decision support to the IT staff of agricultural enterprises, who need to know whether the business process used on their farm complies with the rules currently in force, an activity that is performed by searching for special patterns that summarise international regulations. For this first objective, the specific activities of the BPMN need to be annotated with the generic types of activities defined in the patterns; ii) Provide decision support to Smart Irrigation Systems, which must activate the actuators for automatic irrigation of agricultural fields if and only if a series of weather conditions are fulfilled. For this second objective, it is necessary to populate the domain ontologies with the values of the various sensor readings.

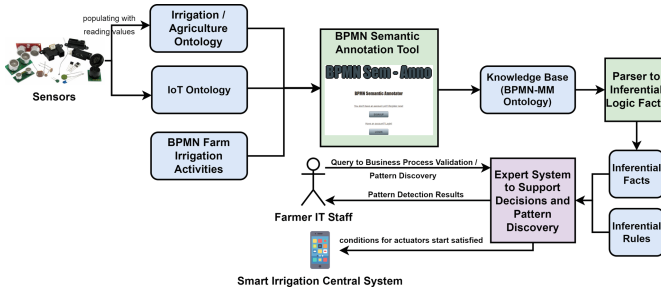


Fig. 1. Proposed methodology.

3.1 The Knowledge Base: An Introduction to the IRRIG Ontology

The **IRRIG ontology**¹ was considered for the implementation of the proposed system. This ontology is written in OWL (Ontology Web Language), and models all the data (measurement data, aggregated data and inferred data from the reasoning process) of the experiments realised in 2019 by **TSCF**², centre **INRAE**³. This experimentation is to evaluate an irrigation system based on the irrigation method Irrinov. Within the IRRIG ontology, created by researchers at the Polytechnic University of Madrid, in collaboration with the University of Clermont Auvergne, is possible to find all the elements needed to assess the state of the irrigation system analysed. Many of these elements were derived from another

¹ <https://irstea.github.io/irrig/OnToology/ontology/irrig.owl/documentation/index-en.html>.

² Technologies et Systèmes d'information pour les agrosystèmes Clermont-Ferrand.

³ National Research Institute for Agriculture, Food and the Environment.

ontology, which is imported into the IRRIG ontology: the extended ontology corresponds to the **CASO (Context-Aware System Observation) ontology**⁴ [9], developed by TSCF, which provides the classes and properties that allow users to model the context-aware system and its observations with a semantic approach. The CASO ontology, extends two standard ontologies: SSN (Semantic Sensor Network)⁵ and SAREF (Smart Applications REference)⁶ [10]. The Semantic Sensor Network (SSN) ontology describes sensors and their observations, the involved procedures, and the studied features of interest. The SSN ontology includes the **SOSA (Sensor, Observation, Sample and Actor)** ontology. The **Smart Applications REference (SAREF) ontology** specifies recurring core concepts in the smart applications domain, the main relationships between these concepts, and axioms to constrain the usage of these concepts and relationships. An overview of the IRRIG ontology and all the relations between the final and extended ontologies are shown in Fig. 2.

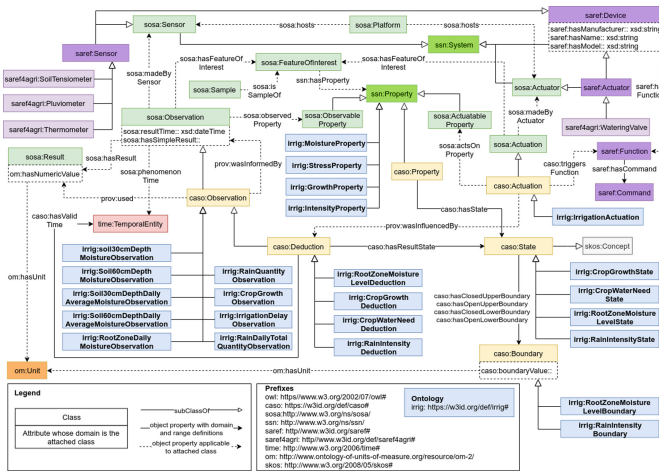


Fig. 2. Overview of the IRRIG ontology.

Due to the nature of the decision-making method, the main role is played by the observations, through which the values measured by the sensors are obtained, which will be analysed, according to the limits imposed by the elements of the **Boundary** class, to arrive at an inference. The **Observation** class has been inherited from the CASO ontology and has been specialised into several subclasses, including, for example, the subclass relating to observations concerning the amount of rain falling on the ground, called **RainQuantityObservation**.

⁴ <https://irstea.github.io/caso/OnToology/ontology/caso.owl/documentation/index-en.html>.

⁵ <http://www.w3.org/ns/ssn/>.

⁶ <https://saref.etsi.org/core/v3.1.1/>.

Within the starting ontology, there were no individuals representing the various observations made, so an ontology population activity was performed, inserting four instants of time in which all the measurements, related observations and, consequently, deductions were made. Thus, the first individuals added correspond to the time instants, which were represented as objects of the class **Temporal entity**. Each observation is associated with a specific time through the object property *phenomenonTime* inherited from the SOSA ontology. In addition to the information on the time instant at which the observation takes place, each observation is associated with information on the kind and value of the sensor reading. Once the observation was added, it was also necessary to create a **Deduction** class relating to the deductions made based on the readings of the recorded observation values.

4 BPMN Analysis, Semantic Annotation and Rules Application

The examined BPMN provides a description of the process of automated irrigation, where an unmanned irrigation system autonomously decides when to provide water to plants, based on the soil moisture and on the recent rainfalls. The proposed BPMN is shown in Fig. 3, and it is the result of an extensive modification of the one proposed in work [6] by the University of Lisbon. The BPMN has three pools: Farmer Pool, Central Pool, IoT Pool. In the farmer pool, two main flows are defined: the first flow deals with the start-up of the entire system, which can only start, however, after validating the tensiometers in the system. Thus, the farmer will have to request the values measured by the tensiometers, after which, once received, he will have to analyse them, and examine their validity and, only in the event of a positive outcome, will he have to start the decision support system. The second flow corresponds to the observation of the growth status of the culture, when required, with the corresponding reporting of the information obtained. In the central pool, all the activities of the irrigation system are defined, in particular, if and only if the farmer has started the system, the system takes care of sending requests to the sensors, analysing the values received and transmitting the signals to the actuators. In the IoT pool, there is a division into two lanes, dedicated to sensors and actuators respectively. In general, the sensors receive a request, read the values and forward them to the sender, while the actuators start their activity after receiving the signal from the central system.

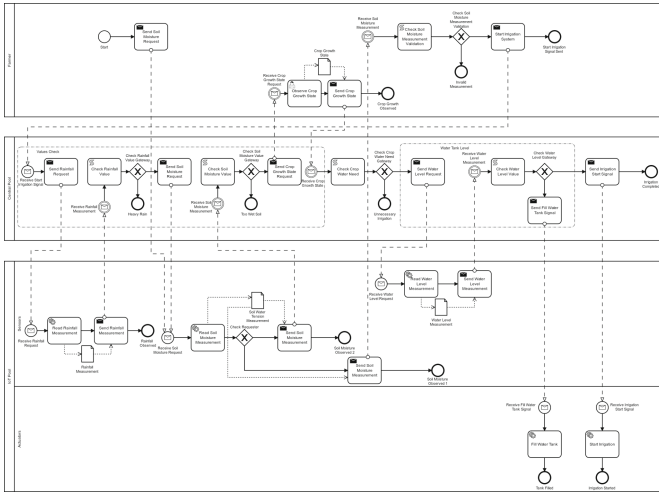


Fig. 3. The BPMN proposed.

4.1 BPMN Semantic Annotation and Pattern Discovery

In order to build the system to support decisions based on inferential rules, the Irrinov method was considered, which in many regions of France is already the most widely used guide for making decisions on the best time to start irrigation, based on fundamental crop and soil properties obtained from measurements with soil moisture sensors and pluviometers. The sensors required to apply this method are: i) Two sets of three tensiometers positioned at two different depths (30 and 60 cm); ii) A pluviometer that measures the water level inside the irrigation tank; iii) A weather station with a second pluviometer to measure the amount of water received by the crop during rainfall. In order to ensure a satisfactory water supply without waste, the method defines the system requirements and the algorithm for decision-making, which were appropriately translated into a set of **Prolog** inference rules. It was necessary to convert the assertions of the OWL knowledge base into a set of Prolog facts using the **Thea parser**. In the Table 1 are shown the main rules that we have extrapolated from the Irrinov method. In these rules, reference is made to the BBCH measurement scale⁷.

Prolog implementation of the rule *Need Water V2/V7* is shown below:

```

v2v7NeedWaterState(Time) :- instant30CmResults(ResultsList30, Time),
instant60CmResults(ResultsList60, Time),
observationsState(StateList, Time)(member(state_crop_growth_v2, StateList);
member(state_crop_growth_v7, StateList)), member(Value30_1, ResultsList30),
member(Value30_2, ResultsList30),
Value30_1notValue30_2, Value30_1 > 50, Value30_2 > 50, !,
member(Value60_1, ResultsList60),
member(Value60_2, ResultsList60), Value60_1notValue60_2, Value60_1 > 30, Value60_2 > 30, !.
    
```

⁷ <https://www.english.arvalisinstitutduvegetal.fr>.

Table 1. Inferential rules implemented in the expert system

Rule	Description
Need Water V2/V7	The system must communicate the need for water for the culture if its growth status is V2 or V7, at least two tensiometers at 30 cm measure values greater than 50cbar and at least two tensiometers at 60 cm measure values greater than 30cbar.
Need Water V7d20/R1	The system must report the need for water for the culture if its growth status is V7d20 or R1, at least two tensiometers at 30 cm measure values greater than 40cbar and at least two tensiometers at 60 cm measure values greater than 20cbar
Irrigation	The system must start irrigation when the water tank is full, the rain intensity is light and the crop needs water.
Contact Problem	The system must report a contact problem between the probe and the ground at a given instant if at least one of the tensiometers results in a value greater than a set limit, rendering the measurements invalid
Acceptable Values	The system should only accept the values resulting from the tensiometers at a given instant if the difference between the average values of the results provided by the tensiometers at different depths is within a predetermined range

Soil irrigation is necessary if and only if the growth status of a crop is V2 or V7, i.e. if it has between 2 and 7 leaves (growth levels are represented using the **IOWA notation of Arvalis**, which takes into account the number of leaves and the reproductive phase of the crop), and the soil moisture is such that at least two of the three tensiometers placed at a depth of 30 cm measure values greater than 50 cbar and at least two of the three tensiometers placed at a depth of 60 cm measure values greater than 30 cbar. Since there are six tensiometers, divided into two sets, located at two different depths in the system under consideration, it was necessary to distinguish the measurements taken by the two sets of sensors, which are returned, given a certain instant of time *Time*, in two different lists thanks to the sub-rules *instant30CmResults* and *instant60CmResults*. Then, the *observationsState* sub-rule returns the growth state of the crop, after which it is checked whether this state is present within the list of states of the elements characterising the plant, using the Prolog built-in *member* predicate. If it is present, the values measured by the two tensiometers are taken from the respective lists by recapturing the member predicate, and then verifying that these values are greater than the lower limit imposed by the rule. A Business Process Pattern has been extrapolated from the Irrinov method, which defines all indispensable activities, which, if carried out in the right order, can ensure that the farms business process complies with current regulations and ensure optimal automatic irrigation. As can be seen from the pattern shown in Fig. 4, the system must verify that irrigation only starts after checking the intensity of the rainfall, the soil moisture, the crops need for water, and the water level inside

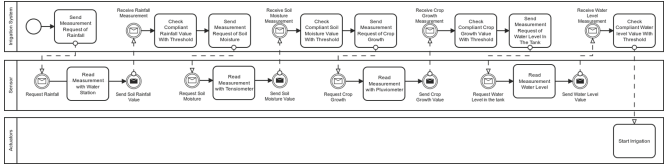


Fig. 4. Business process pattern extrapolated from the Irrinov method.

the tank using the appropriate sensors. If this series of measurements respects the thresholds imposed by the Irrinov method, then the system can start the irrigation actuators. To verify whether such a pattern is present in the business process, manual semantic annotation is performed, in which the various tasks of the BPMN were associated with ontology concepts modelling the pattern activities.

An example of an inserted annotation is as follows: the BPMN task *Send Rainfall Request* was annotated with *SIT* individual of the ontology using *has_performerLink* annotation, to indicate that this activity was performed by the Smart Irrigation System, but also with *RainfallRequest* class of the ontology using *activity_is_a_kind_of* annotation, to add the appropriate semantics to the BPMN to map the elements of the BPMN with the generic activities of the business process pattern. The pattern search in the BPMN is carried out using a Prolog rule which performs two checks: i) there must exist at least one path in the BPMN containing at least one task for each type of activity present in the pattern, or in other words, at least one task or event in the BPMN annotated semantically with one ontology concept modelling the activity and one modelling the actor performing the activity; ii) the pattern activities identified in the path must respect the same causal order defined in the pattern. To do this, given two elements of the BPMN semantically annotated with pattern activities, there must be a path from the first element to the second element. If these conditions are satisfied, then the BPMN respects the pattern and thus is compliant with the regulations imposed by the Irrinov method.

5 Conclusion and Future Works

In this paper, a methodology for the smart management of irrigation systems using a semantic-based approach is proposed. The methodology uses BPMN semantic annotation and logical inference techniques on an OWL knowledge base made up of a domain ontology of IoT and agricultural concepts, and an ontology describing all structural aspects of a BPMN concerning the automatic irrigation procedure that a Smart Irrigation System can perform by conducting checks based on measurements taken by sensors. In the future, the intention is to extend this methodology to other aspects of Smart Agriculture, to provide the IT staff of agri-businesses with a framework to support them in managing their activities more effectively.

Acknowledgments. The work described in this paper has been supported by the Project VALERE SSCeGov - Semantic, Secure and Law Compliant e-Government Processes.

References

1. Ayaz, M., Ammad-Uddin, M., Sharif, Z., Mansour, A., Aggoune, E.-H.M.: Internet-of-things (iot)-based smart agriculture: toward making the fields talk. *IEEE Access* **7**, 129551–129583 (2019)
2. Bouthier, A., Deumier, J.M., Lacroix, B., et al.: IRRINOV, a farmer-oriented scheduling method for maize, cereals and pea irrigation. In: Improved irrigation technologies and methods: Research, development and testing. Proceedings ICID International workshop, Montpellier, France, 14-19 September 2003, pp. 1–9. International Commission on Irrigation and Drainage (ICID) (2003)
3. Di Martino, B., Cascone, D., Colucci Cante, L., Esposito, A.: Semantic representation and rule based patterns discovery and verification in eProcurement business processes for eGovernment. In: Barolli, L., Yim, K., Enokido, T. (eds.) *CISIS 2021*. LNNS, vol. 278, pp. 667–676. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79725-6_67
4. Di Martino, B., Graziano, M., Colucci Cante, L., Esposito, A., Epifania, M.: Application of business process semantic annotation techniques to perform pattern recognition activities applied to the generalized civic access. In: Barolli, L. (ed.) *CISIS 2022*. Lecture Notes in Networks and Systems, vol. 497, pp. 404–413. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08812-4_39
5. Di Martino, B., Graziano, M., Colucci Cante, L., Ferretti, G., De Oto, V.: A semantic representation for public calls domain and procedure: housing policies of Campania Region case study. In: Barolli, L. (ed.) *CISIS 2022*. Lecture Notes in Networks and Systems, vol. 497, pp. 414–424. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08812-4_40
6. Domingos, D., Respício, A., Martins, F., Melo, B.: Automatic decomposition of IoT aware business processes-a pattern approach. *Procedia Comput. Sci.* **164**, 313–320 (2019)
7. Friha, O., Ferrag, M.A., Shu, L., Maglaras, L., Wang, X.: Internet of things for the future of smart agriculture: a comprehensive survey of emerging technologies. *IEEE CAA J. Autom. Sinica* **8**(4), 718–752 (2021)
8. Khatoon, P.S., Ahmed, M.: Importance of semantic interoperability in smart agriculture systems. *Trans. Emerg. Telecommun. Technol.* **33**(5), e4448 (2022)
9. Nguyen, Q.-D., Roussey, C., Poveda-Villalón, M., de Vaulx, C., Chanet, J.-P.: Development experience of a context-aware system for smart irrigation using CASO and IRRIG ontologies. *Appl. Sci.* **10**(5), 1803 (2020)
10. Poveda-Villalón, M., Nguyen, Q.D., Roussey, C., de Vaulx, C., Chanet, J.P.: Ontological requirement specification for smart irrigation systems: a SOSA/SSN and SAREF comparison. In: 9th International Semantic Sensor Networks Workshop (SSN 2018), vol. 2213, p. 16, Monterey, United States, October 2018. CEUR Workshop Proceedings

11. Rak, M., Granata, D., Di Martino, B., Colucci Cante, L.: A semantic methodology for security controls verification in public administration business processes. In: Barolli, L. (ed.) CISIS 2022. Lecture Notes in Networks and Systems, vol. 497, pp. 456–466. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08812-4_44
12. Rotondi, D., Straniero, L., Saltarella, M., Balducci, F., Impedovo, D., Pirlo, G.: Semantics for wastewater reuse in agriculture. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 598–603 (2019)



Applying CI/CD Process to Improve the Speed and Critical Quality of Perfective Maintenance

Sen-Tarng Lai^{1(✉)} and Fang-Yie Leu²

¹ Department of Information Technology and Management, Shih Chien University, Taipei 10462, Taiwan

stlai@g2.usc.edu.tw

² Department of Computer Science, Tunghai University, Taichung 40704, Taiwan

leufy@thu.edu.tw

Abstract. Information system is an essential tool for survival in every industry, and it is also the key to enhancing market competitiveness. In order to extend the life cycle, the Information systems must under maintenance status to meet the continuous requirements of users and cope with the environment changes. Perfective maintenance is the most frequent and the most critical task for the growing enterprises and organizations. Therefore, improving the speed and workflow quality of perfective maintenance is a necessary condition to increase market competitiveness. Agile software development and DevOps use CI/CD process to improve the efficiency and quality of software development, operations and maintenance. In this paper, we define major quality items from the CI/CD process and propose a process quality measurement model to quantify and improve process quality. Applying a high quality CI/CD process, the speed and quality of perfective maintenance are sufficient to meet user requirements and environmental changes.

1 Introduction

In the age of information and the internet, in order to achieve sustainable operation, enterprises and organizations must have a perfect and continuously evolving information system. After the information system is delivered to the owner, it will enter the maintenance phase [1]. With the changes in the environment and the customers constantly putting forward new requirements for the system, in order to maintain the normal operation of the information system and meet the needs of the environment and customers, it must continue to grow under the maintenance state. Among the type of software maintenance, perfective maintenance requests have the highest frequency [2]. For growing businesses and organizations, perfective maintenance has the greatest impact on the market competitiveness and operational efficiency of the entire business organization. How to effectively improve the speed and workflow quality of the perfective maintenance of information systems is the key to the sustainable survival and growth of enterprises and organizations.

The current software development methods and technologies attach great importance to the interaction and communication between staff and between staff and users. Agile software development proposed by 17 software practitioners in 2001 [3], adopts the

iterative and incremental development (IID) method, and attaches great importance to the interaction between developers, end users, product owners, and stakeholders. [4] Replacing complete development documents and standard process specifications, and being able to adapt to changes and adjustments as requested by customers at any time, has become a development method that is currently accepted by most enterprises and organizations. Among them, the Continuous Integration/Continuous Delivery (CI/CD) process [5] greatly improves the speed and quality of product delivery, and can quickly meet the needs of customers. In addition, DevOps proposed in 2009 is a development method that combines the development and operation cooperation of developers with the culture of operators, which reduces the personnel conflict between the two departments and effectively improves the benefit of cooperation [6, 7]. In DevOps, the CI/CD process speeds up product maintenance and delivery time, greatly reduces customer waiting time, and effectively enhances the market competitiveness of customers, enterprises and organizations.

A high-quality CI/CD process can speed up the tasks of perfective maintenance and effectively improve customer competitiveness. This paper collects major quality items of the CI/CD process and proposes the CI/CD Process Quality Measurement (CPQM) model to identify the quality defects of the CI/CD process, and develop specific improvement methods. This enables perfective maintenance operations can complete rapid deployment and maintain high product quality. In addition, enterprises and organizations can continuously undertake new customer requirements and changing environments. Section 2 discusses the importance and challenges of perfective maintenance in the enterprise or organization. Section 3 discusses the steps of CI/CD process and the major quality items. Section 4 based on the major quality items, proposes a process quality measurement (CPQM) model to quantify and improve the operation quality of CI/CD process. Section 5 evaluates the advantages of perfective maintenance with CI/CD process. Section 6 describes the importance of a high-quality CI/CD process to expedite perfective maintenance and rapid deployment. This paper proposes the CPQM model to quantify and improve the operation quality of the CI/CD process to overcome three challenges of perfective maintenance.

2 The Importance and Challenges of Perfective Maintenance

Perfective maintenance is an important type of software maintenance that must face the challenges of rapidly deploying and maintaining high product quality.

2.1 Importance of Perfective Maintenance

In order to extend the life cycle of the information systems, the delivered systems must enter the maintenance phase, and constantly accept the new functional requirements, changing requirements, and environmental evolution. There are four types of software maintenance [2] described as follows:

- **Corrective Maintenance:** It mainly occurs in the early stage of system delivery. There are still many residual errors and defects in the information system. The end user submits a Corrective Maintenance request for the errors and defects in the information system.
- **Perfective Maintenance:** Extend, change or adjust the existing requirement items of the information system. This type has the highest effort (about 50%) in the maintenance phase [2].
- **Adaptive Maintenance:** After the information system operates for a period of time, in order to keep pace with the times, the configured software or hardware facilities must be updated. At this time, the maintenance staff puts forward an Adaptive Maintenance request.
- **Preventive Maintenance:** In order to maintain the normal operation of the information system, preventive measures must be taken in advance before abnormal conditions occur, such as patching security holes, expanding storage space, and replacing hardware configure.

The characteristic differences between the four software maintenance types are shown in Table 1.

Table 1. Attributes Comparison Table for four software maintenance types

Maintenance Type	Occurred period	Effort*	Impacts
Corrective Maintenance	After delivery	21%	Normal operations
Perfective Maintenance:	Any time	50%	Competitiveness, user confidence
Adaptive Maintenance	After a period of time	25%	Performance, efficiency
Preventive Maintenance	Specific period	4%	Security, efficiency, reliability

Perfective maintenance refers to the process of modifying software or applications to achieve new requirements involving functional adjustments or extensions. In the software maintenance phase, perfective maintenance request accounts for the highest effort (about 50%) [2], which has a high influence on growing enterprises and organizations.

2.2 Challenges of Perfective Maintenance

In the age of information and network, people's daily activities rely more and more on information systems. In order to improve the service quality of enterprises and organizations and meet the various needs of users, the information system should support convenient and uninterrupted services to effectively reduce interrupted maintenance times. In addition of shorten the maintenance time and nonstop services, (Information System) Perfective maintenance also should maintain the product's critical quality. Perfective maintenance should over three major challenges:

- The deployment must be completed quickly to meet user requirements.
- System maintenance as possible without suspending services, reducing the inconvenience of users.
- Must maintain product critical quality (security, correctness, integrity, consistency) to enhance users' trust.

In the process of perfective maintenance, the operation mode of testing and integration tasks is the most critical. In addition to impacting the speed and effect of product delivery and deployment tasks, the critical quality of products will also be affected by this operation. Therefore, perfective maintenance must formulate a set of suitable and high-quality operation procedures to overcome the three challenges of rapid deployment, reducing user inconvenience, and maintaining high product quality. In order to continuously deliver information systems that meet the needs of enterprises and organizations to enhance their market competitiveness, agile software development and DevOps adopt an automation-based CI/CD process and interaction and communication, which can speed up product deployment and still focus on product important quality, making CI/CD process has become a powerful tool for agile software development and DevOps success [7, 8]. Perfective maintenance is consistent with the goals of agile software development and DevOps, and should make use of the advantages of the CI/CD process to overcome three challenges and meet user requirements.

3 CI/CD Operation Process and Major Quality Items

3.1 CI/CD Process Workflow

The CI/CD process is the continuous development, testing, delivery, and rapid deployment of high product quality, the ideal combination for perfective maintenance of frequent requirements [9]. Combining multiple software tools for development, testing, delivery, and deployment is basic to the automation of the CI/CD pipeline process [10]. Crucially, the CI/CD process can be subdivided into four core steps (shown in Fig. 1), in which CI missions are divided into two steps of affected items identification and automatic testing and integration:

- Step 1. Affected items Identification: According to the maintenance request, apply the version control system and configuration management system to quickly identify affected related software items from the software repository.
- Step 2. Automatic testing and integration: Affected items have to pass unit testing, integration testing, regression testing, and functional testing. Using automatic test tools to reduce staff participation can speed up the efficiency and quality of testing and integration.

CD process mission includes two step process for continuous delivery and continuous deployment:

- Step 3. High qualified delivery: Convert changing requirements into acceptance criteria, use tools to assist system testing, acceptance testing, and delivery procedure, confirm that the new version has incorporated key qualities such as correctness, completeness, consistency, and security. And deliver procedures facilitate the rapid and continuous delivery of new releases.
- Step 4. Rapid deployment: Convert the operating environment requirements into deployment setting parameters, and then use installation assistance tools and deployment procedures to quickly complete the deployment of the new version.

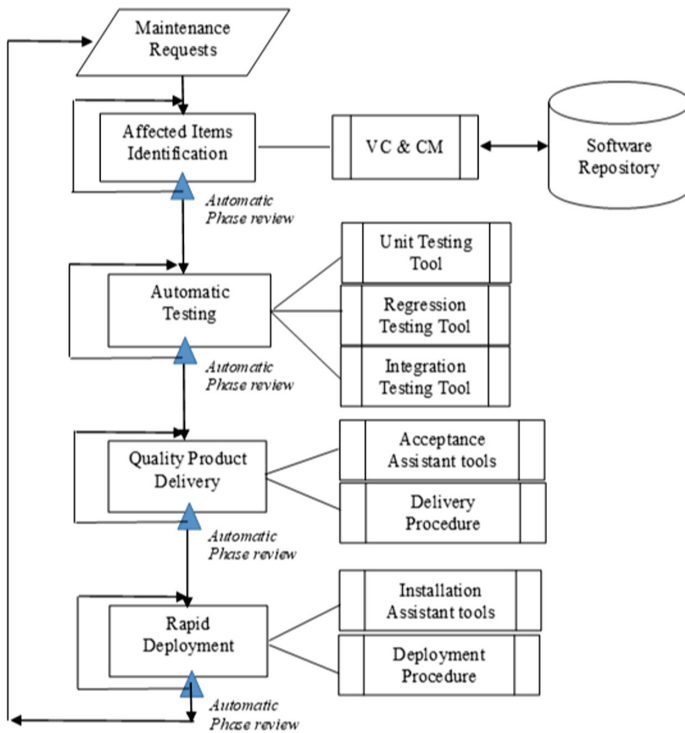


Fig. 1. CI/CD process operation flow

3.2 Quality Items of CI/CD Process

The CI/CD process is an automated toolchain to improve the operational efficiency and product quality of the Agile Software Development and DevOps environment. The precondition of automation is that a set of standards must be established for the input and output formats of each stage of the CI/CD process so that the automation process can greatly reduce human intervention, reduce error rates, and improve process efficiency and product quality. To achieve rapid maintenance and deployment tasks, a CI/CD pipeline

should have steps such as Identification, Automated testing, quality product delivery, and rapid deployment (shown as Fig. 1), in addition to automation capabilities, each stage must be highly integrated. The quality items that should be provided in each stage are described as follows:

- **Affected Items Identification:** In this phase, for quickly and correctly identifying the affected item of maintenance request, the software repository (SR), configuration management (CM) system, and version control (VC) system quality items should be considered.
- **Automatic Test:** Software testing is a complex mission and always takes much time. Applying testing tools can greatly reduce staff participation and speed up the efficiency and quality of software testing and integration. Therefore, in this phase, the automated unit testing, integration testing, regression testing, and functional testing quality items should be considered.
- **Quality Product Delivery:** Information systems with high quality and security is an important factor to get user confidence. Convert maintenance requirements into acceptance criteria, use acceptance assistance tools for system testing, acceptance testing, and training courses, and confirm that the new version has incorporated critical qualities such as correctness, integrity, consistency, and security [11]. In addition, the quality product delivery procedure assists the rapid and continuous delivery of new releases. In this phase, the acceptance assistance tools and quality delivery procedure quality items should be considered.
- **Rapid Deployment:** Convert the operating environment requirements into deployment setting parameters, and then use deployment assistance tools and deployment procedure to quickly complete the deployment of the new version. In this phase, the installation assistance tools and rapid deployment procedure quality items should be considered.

4 CD Process Measurement Model and Improvement Manner

4.1 Quantified Measurement Model

For quantifying and improving the CI/CD process, we collected the major quality items of CI/CD process. And, based on the linear combination model [12], proposes a CPQM model that combines AII, AT, QPD, and RD four quality measurements. Inspect and review factors of quality items, software professionals and experienced maintainers can draw up quantitative values of basic quality items. The quantified value approaching 1 represents good quality, and a quantified value approaching 0 represents poor quality. In the model, the senior software engineers assign the weighted value between 0 and 1. The weighted value close to 1 indicates that the quality item is important for quality measurement. The CPQM model describes as follows (shown in Fig. 2):

- (1) **Affected items identification:** For quickly identifying the affected items from the software repository, the Configuration Management (CM), Version Control (VC), and Software Repository (SR) should have high quality. For this, Affected Items

Identification Measurement (AIIM) combines CM, VC, and SR three quality items. Combination formula is shown as Eq. (1):

$$\begin{aligned}
 & \text{AIIM: Affected Items Identification Measurement} \\
 & \text{CMQ: Configuration Management Quality} \quad W_1: \text{Weight of CMQ} \\
 & \text{VCQ: Version Control Quality} \quad W_2: \text{Weight of VCQ} \\
 & \text{SRQ: Software Repository Quality} \quad W_3: \text{Weight of SRQ} \\
 & \text{AIIM} = W_1 * \text{CMQ} + W_2 * \text{VCQ} + W_3 * \text{SRQ} \quad W_1 + W_2 + W_3 = 1
 \end{aligned} \tag{1}$$

- (2) Automatic test quality needs to consider the combination quality of unit testing (UT), regression testing (RT), and integration testing (IT) three automatic tools. For this, Automatic Test Measurement (ATM) combines UT, RT, and IT three quality items. Combination formula is shown as Eq. (2):

$$\begin{aligned}
 & \text{ATM: Automatic Test Measurement} \\
 & \text{UTQ: Unit Testing Quality} \quad W_1: \text{Weight of UTQ} \\
 & \text{RTQ: Regression Testing Quality} \quad W_2: \text{Weight of RTQ} \\
 & \text{ITQ: Integration Testing Quality} \quad W_3: \text{Weight of ITQ} \\
 & \text{ATQM} = W_1 * \text{UTQ} + W_2 * \text{RTQ} + W_3 * \text{ITQ} \quad W_1 + W_2 + W_3 = 1
 \end{aligned} \tag{2}$$

- (3) Quality Product Delivery Measurement (QPDM) should consider acceptance assistance tools (AAT) and quality product delivery procedure (QPDP) quality items to confirm the new version has high quality and can rapid delivery. For this, QPDM combines AAT and QPDP two quality items. Combination formula is shown as Eq. (3):

$$\begin{aligned}
 & \text{QPDM: Quality Product Delivery Measurement} \\
 & \text{AATQ: AAT Quality} \quad W_1: \text{Weight of AATQ} \\
 & \text{QPDPQ: QPDP Quality} \quad W_2: \text{Weight of QPDPQ} \\
 & \text{QPDM} = W_1 * \text{AATQ} + W_2 * \text{QPDPQ} \quad W_1 + W_2 = 1
 \end{aligned} \tag{3}$$

- (4) Rapid Deployment Measurement (RDM) needs to consider the quality of installation assistance tools (IAT) and rapid deployment procedure (RDP). For this, RDM combines IAT and RDP two quality items. Combination formula is shown as Eq. (4):

$$\begin{aligned}
 & \text{RDM: Rapid Deployment Measurement} \\
 & \text{IATQ: Installation assistant Tools Quality} \quad W_1: \text{Weight of IATQ} \\
 & \text{RDPQ: Rapid Deployment Procedure Quality} \quad W_2: \text{Weight of RDPQ} \\
 & \text{RDM} = W_1 * \text{IATQ} + W_2 * \text{RDPQ} \quad W_1 + W_2 = 1
 \end{aligned} \tag{4}$$

- (5) CI/CD Process Measurement (CIDM) combines four phase quality measurements which include AIIM, ATM, QPDM, and RDM. And, according to the influence of quality measurement, the weighted value of linear combination equations is assigned. Combination formula is shown as Eq. (5):

$$\begin{aligned}
 & \text{CIDM: CI/CD Process Measurement} \\
 & \text{AIIM: Affected Items Identification Metric} \quad W_1: \text{Weight of AIIM} \\
 & \text{ATM: Automatic Testing Metric} \quad W_2: \text{Weight of ATM} \\
 & \text{QPDM: Quality Delivery Metric} \quad W_3: \text{Weight of QPDM} \\
 & \text{RDM: Rapid Deployment Metric} \quad W_4: \text{Weight of RDM} \\
 & \text{CIDM} = W_1 * \text{AIIM} + W_2 * \text{ATM} + W_3 * \text{QPDM} + W_4 * \text{RDM} \\
 & \quad W_1 + W_2 + W_3 + W_4 = 1
 \end{aligned} \tag{5}$$

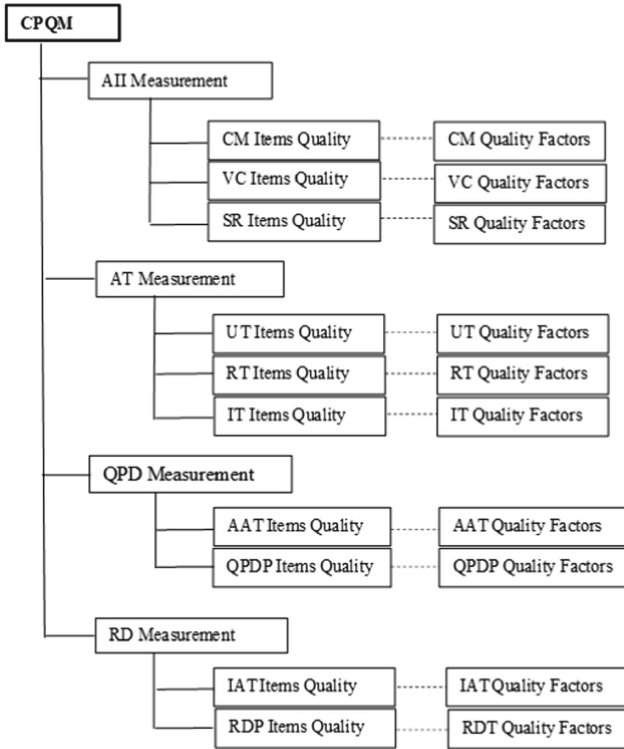


Fig. 2. Architecture of CPQM model

4.2 CI/CD Process Improvement Measures

Quality quantification can help identify the item quality defects of the CI/CD process and then take appropriate improvement measures. The following describes the process quality improvement measures based on the CPQM model:

IF $CIDM < 0.5$, according to Eq. (5), the defect of quality measurements should be identified, and apply the rule-based improvement mechanism:

The template improvement rule of AII measurement shows as follows:

IF the AII measurement $< 0.5*$ (*threshold can be adjusted)

THEN according to Eq. (1), check the relevant items of VC, CM, and SR one by one and determine the poor quality items, and list the concerned quality factors that need to be improved.

AT measurement, QPD measurement, and RD measurement can instance the improvement rules using the template improvement rule.

5 Evaluation of CI/CD Process of Perfective Maintenance

The traditional software maintenance work is always performed by personnel who lack software development experience. Under the condition that incorrect, incomplete, and inconsistent development resources management system, the workload of personnel is heavy, the pressure is high, and the morale is low. Software maintenance is not easy to grasp timeliness and expected results. Recently software maintenance has been enhanced the automatic tools to increase productivity and quality. However, perfective maintenance always can't meet the user instantly requirements. In this paper, we apply the automated CI/CD process to perfective maintenance for improving rapid deployment and high product quality. This section discusses and evaluates the advantages of the CI/CD process of perfective maintenance from three aspects (shown in Table 2):

1. Traditional perfective maintenance: lack of automatic operations flow, for reducing the maintenance cost and frequency, several maintenance requests will be merged into one time. It should reduce the maintenance manpower, however, it makes more inconvenience and fully can't meet the instant requirements of the user.
2. Perfective maintenance with automatic tools system: deployment speed and high product quality of perfective maintenance cannot make sure the rapid deployment requirements of the user. Lack of tools integration, quality delivery, and improvement mechanism, perfective maintenance process can't reach rapid deployment and required high quality.
3. Perfective maintenance with CI/CD process: Based on automatic tools and improvement mechanisms, the CI/CD process can meet three challenges of perfective maintenance: rapid deployment, concern the critical quality, and reduced user inconvenience.

Table 2. CI/CD Process evaluation in perfective maintenance

Maintenance styles	Traditional	Automation tools system	Applying CI/CD process
Characteristics			
1. Software Repository Management System	Weak	Strong	Strong
2. Automatic Testing	General	Strong	Strong
3. Quality Product Delivery	Weak	General	Strong
4. Rapid Deployment	Weak	General	Strong

6 Conclusion

Quickly meeting various requirements is an important factor to improve the service quality of information systems. With the continuous changes in the environment, enterprises and organizations must constantly adjust and add service items that keep pace with the times to meet user requirements. For growing enterprises and organizations, perfective maintenance has a great influence on market competitiveness and operational effectiveness. How to effectively improve the delivery speed and high product quality of perfective maintenance is the key to the sustainable survival and growth of enterprises and organizations. Agile software development and DevOps are widely used software development, operation, and maintenance methods. Both methods use CI/CD process to rapidly deploy and maintain high product quality. This paper collects the quality items of the CI/CD process, and proposes the CPQM model for evaluating and improving the quality of the CI/CD process. Perfective maintenance applies high quality CI/CD process that can overcome the three challenges of rapid deployment, reducing user inconvenience, and maintaining high product quality. Responding to the ever-changing environment and new requirements of customers, helping enterprises and organizations to improve their market competitiveness.

References

- Schach, S.R.: Object-Oriented Software Engineering, vol. 7. McGraw-Hill, New York (2008)
- Aggarwal, K.K.: Software Engineering. New Age International (2005)
- Beck, K., et al.: Manifesto for Agile Software Development (2001). <http://www.agilemanifesto.org/>
- Larman, C., Basili, V.R.: Iterative and Incremental Development: A Brief History. Computer, p. 48. IEEE CS Press (2004)
- Schwaber, K., Beedle, M.: Agile Software Development with Scrum, vol. 1. Prentice Hall, Upper Saddle River (2002)
- Ebert, C., Gallardo, G., Hernantes, J., Serrano, N.: DevOps. *IEEE Softw.* **33**(3), 94–100 (2016)
- Leite, L., Rocha, C., Kon, F., Milojevic, D., Meirelles, P.: A survey of DevOps concepts and challenges. *ACM Comput. Surv. (CSUR)* **52**(6), 1–35 (2019)
- Shahin, M., Babar, M.A., Zhu, L.: Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access* **5**, 3909–3943 (2017)

9. Purohit, K.: Executing DevOps & CI/CD reduce in manual dependency. *IJSDR* **5**(6), 511–515 (2020)
10. Jackson, L.: The CI/CD pipeline. In: *The Complete ASP.NET Core 3 API Tutorial*, pp. 305–347. Apress, Berkeley (2020)
11. Hu, P., Chaowen, C., Ma, Y., Wang, X.: Acceptance testing optimization method for continuous delivery. In: *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT)*, pp. 168–173. IEEE (2021)
12. Fenton, N.E.: *Software Metrics - A Rigorous Approach*. Chapman & Hall (1991)



Generating Personalized Phishing Emails for Social Engineering Training Based on Neural Language Models

Shih-Wei Guo¹, Tzu-Chi Chen¹, Hui-Juan Wang¹, Fang-Yie Leu²(✉),
and Yao-Chung Fan¹

¹ National Chung Hsing University, Taichung, Taiwan

² TungHai University, Taichung, Taiwan

leufy@thu.edu.tw

Abstract. To prevent phishing attacks, social engineering training is a practical way by reinforcing the concepts of being aware of phishing emails. However, existing social engineering training relies on manual planning and artificially design, which suffers from scale and cost concerns. In this paper, we explore the idea of using natural language generation techniques for automatic social engineering training phishing mail generation. We present AI-Phishing, a novel phishing mail generation to facilitate personalized social engineering training planning. Users can utilize AI-Phishing to generate personalized phishing emails according to a given title and keywords.

1 Introduction

Phishing attacks bring significant concerns for enterprise security. Holding social engineering training [10] to being aware of phishing attacks is a practical way for security protection [5]. However, existing social engineering training relies on manual planning and artificially design, which suffers from scale and cost concerns.

In this paper, we present *AI-Phishing*, a novel Phishing SET curation system based on natural language generation (NLG) techniques. The core of AI-Phishing is its text generation module which generates mail content based on given keywords, as illustrated in Fig 1.

The design goal of AI-Phishing is to facilitate enterprise social engineering training. Users can utilize AI-Phishing to generate personalized phishing according to the given title and keywords. Personalized content can be generated for employees based on their expertise, interests, or personal profiles. This paper reports our AI-Phishing design and investigate the following questions:

- How to fine-tune pre-trained language models based on email corpus?
- How to boost the model quality by considering mail content diversity and keeping up with news events?
- What is the quality and feasibility of the auto-generated emails?

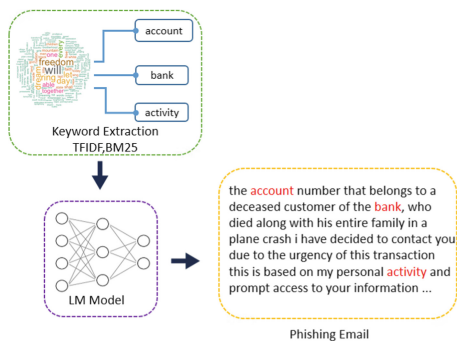


Fig. 1. Automatically generating mails based on user-indicated keywords.

2 Methodology

2.1 Training Email Generation Model

In this subsection, the training setting for the email text generation module is presented.

The email text generation module consists of components as depicted in Fig 2. First, a language model component takes charge of basic text generation capabilities. Second, a keyword extractor component (such as TF-IDF or BM25) takes charge of extracting keywords from the email text. The extracted keywords from an email text are served as the training data for conditional email generation learning.

Specifically, during the training phase, for a given email training instance with content body B with (1) keywords $\{k_1, \dots, k_n\}$ extracted by the keyword extractor and (2) email subject S , the input to our model is formulated as follows.

$$\mathbb{M}(k_1, \dots, k_n, [\text{SEP}], S) \rightarrow B \quad (1)$$

The learning objective is to generate B based on a given subject and keywords. The goal of indicating the subject is to control the structure of the whole email content and the goal of indicating keywords is to guide the details of the generated mail.

To better see the input, we show a real example in Table 1, where the keywords are set to “COVID-19”, “case”, “hospital” and “CDC” and “Please Reply to Me Soon” is the subject as input and the result shown in the bottom of the table is the target output.

Table 1. A generation result example

Subject:
Please Reply to Me Soon

Keywords:
covid19, case, hospital, CDC

Outputs:
Please help me spread this idea... it's a very good case for the **CDC**. I've been working closely with a number of people to help me understand the **hospitalization** situation and plan for the potential for a **COVID-19** vaccine.

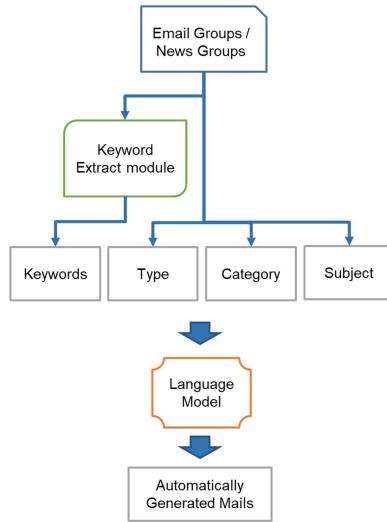


Fig. 2. Neural-phishing-flow-diagram.

2.2 Boosting Performance by Multi-field Learning

By the experiments, we find that the model introduced in the previous subsection suffers from the following two limitations.

- **Candidate Similarity** First, the content generated by the model does not vary much; for the same keywords, the model generates almost the same content. However, for practical usage, we would like to have various text generation results for a better social engineering training curation.
- **Content Diversity** Furthermore, when conducting social engineering training curation, generating text with popular words (such as presidential election, epidemic diseases, or stock market) is critical to the quality of the generated emails.

Therefore, for improving the content diversity, we follow the multi-field learning proposed by [14]. We consider two kinds of fields: (1) mail type (fraudulent and non-fraudulent) field T and (2) content type (news and emails) field C .

For employing the multi-field learning, we use CLAIR collection of fraud email [12] as base email corpus for training T field and use news corpus from [6] and [2] for training C field.

In the CLAIR collection, the email instance is annotated as fraudulent or non-fraudulent. We use the email type as the T field selection for multi-field learning.

Furthermore, for C field, we select business, finance, and COVID19 category’s corpus from [6] and [2] for training C field.

For an indicated email subject S , keyword sets $\{k_1, k_2, k_3, \dots, k_{|n|}\}$ and email type T and category C , we train a generation model $M()$ take T , C , S and the keywords as input to generate the ground truth email content B as output. We formulate the input to the language model as follows.

$$\mathbb{M}(T, [\text{SEP}], C, [\text{SEP}], k_1, \dots, k_n, [\text{SEP}], S) \rightarrow B \quad (2)$$

3 Performance Evaluation

3.1 Dataset Collection

- **Fraud Email Dataset** [12]: We use data set from the CLAIR collection of fraud email. There are 11929 emails with 5187 fraud emails and the rest non-fraud emails.
- **New Category Dataset:** [6] This dataset contains about 200,000 news items from 2012 to 2018 obtained from HuffPost. More than 40 news category tags are provided.
- **COVID19 Fake News:** [2] The data set comes from social media such as Twitter, Facebook, Instagram, etc., and is collected by the CONSTRAINT-2021 shared task, focusing on the task of detecting fake news related to COVID19.

We collected 10k training articles which Email is 1.2k News is 8k, and 2k testing articles which Email is 0.25k and News is 0.75k.

During these statistics, we can conclude that our dataset is more suitable for the generation of phishing emails. First, our dataset has many corpora for finance topics and trend words from emails, and news datasets. This can be observed in Fig 3, The Email corpus contains the country’s location name and government relationship except for finance, The news corpus contains COVID19 and Trump, which are usefully trending words. In other words, that corpus for our task is helped.



Fig. 3. Word cloud for Emails and News entities in the our dataset. Note that basic stopwords are excluded for the word cloud for articles.

3.2 Evaluation

Training Setting. Our models are trained on two NVIDIA®Geforce 1080 Ti™ with a memory of 32 GB. We employ the GPT2 [8] and BART model [3] released by huggingface [13]. The models are with the maximum input length setting to 128 tokens. The AdamW optimizer is applied with an initial learning rate of $5e-4$. All models are set to run 10 epochs. GPT2 has base version (parameter 117 M, 12-layer) and medium version (parameter 335 M, 24-layer), Bart also has base version (parameter 139 M, 12-layer) and large version (parameter 406 M, 24-layer). According to previous studies, models with more parameters usually perform better in text generation tasks. Thus, we also evaluate the performance of GPT2 and Bart of various sizes.

Evaluation Criteria. We conduct performance evaluation by the following criteria:

- **Controlability:** How well does the model generate emails based on given keywords?
- **Diversity** How diverse is the model-generated content?
- **Ability to Pass Spam Filters** How well does the auto-generated email pass the mail spam filter?

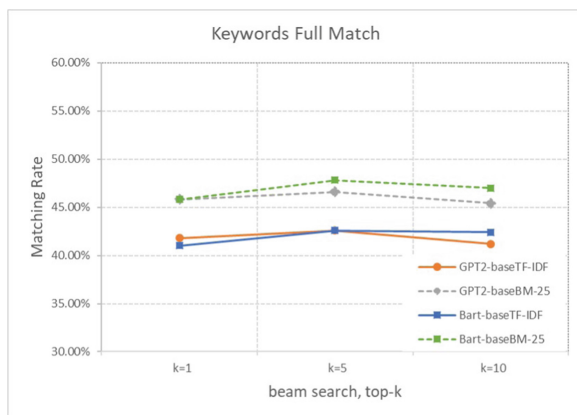


Fig. 4. Controllability comparison on full match rate (base vs base).

3.3 Controllability Evaluation

For evaluating controllability, we use the match rate (the extent that the indicated keywords were presented in the generated contents). We use two-match rate variants (*Partial Match Rate*) and (*Full Match Rate*). In partial match rate, a hit is that one of the indicated keywords is presented in the generated result. On the other hand, in full match rate, a hit is that all indicated keywords are presented in the generated result.

We show the evaluation results of BART and GPT2 in Fig 4 and Fig 5 . In the compared models, we also evaluate the performance by the BM25 [1,9] and TFIDF [11] keyword extractor and the effect of varying the BEAM search parameter k by setting $k = 1$, $k = 5$, and $k = 10$. From the result, we see that the GPT2 model with $k = 10$ and BM25 keyword extractor show a 92% partial match rate and the Bart model with BM25 keywords achieves a 96% match rate. In addition, if the full match rate is considered, it can be found in Fig 4 show that GPT2 with BM25 has 46.61% in the performance of $k = 5$ keywords, and the performance of Bart model with BM25 keywords can reach a score of 48.81%. As can be seen from the above, Bart’s keyword control ability has better performance than GPT2.

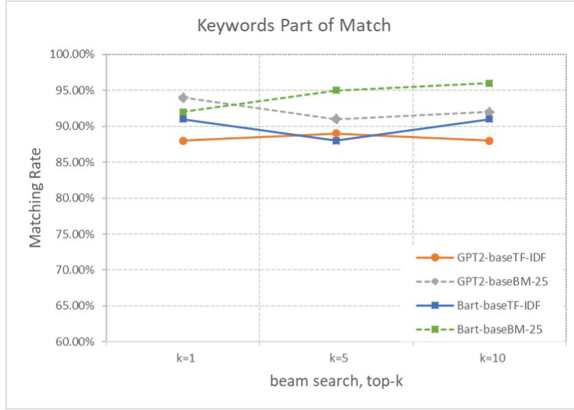


Fig. 5. Controllability comparison on partial match rate (base vs base).

3.4 Diversity Evaluation

In this subsection, we examine the quality and variety of email texts generated by the model. Similarly, we evaluate both models, GPT2 and Bart, using different keyword extraction techniques and using different Beam Search settings.

To verify the performance, we use BLEU [7], Self-BLEU, and ROUGE-L [4]. Specifically, BLEU is used to measure word-level similarity over ground truth and ROUGE-L is for sentence-level evaluation. Also, we use self-BLEU to measure the similarity between generated contents as a diversity index.

We summarize the results in Table 2. We can see that there is no significant difference between BM25 and TFIDF presenting on GPT2-base performance. However, in Table 3, we can see that the performance of keywords extracted by BM25 has been significantly improved when top-k = 10. As long as enough key information and themes are given, the Bart model can be generated to achieve good quality. Also, we can see that diversity (in terms of Self-BLEU [15] to measure repetitions at a distributional level across the whole set of generated samples) also performs well.

Table 2. Performance of GPT2-base.

Performance			
Model	BLEU \uparrow	Self-BLEU \downarrow	Rouge-L \uparrow
GPT-2 (base) TFIDF			
Top-k			
k = 1	0.2083	0.9000	0.3058
k = 5	0.2132	0.6397	0.3071
k = 10	0.2089	0.6590	0.2998
GPT-2 (base) BM25			
Top-k			
k = 1	0.2145	0.9000	0.3159
k = 5	0.2137	0.6151	0.3190
k = 10	0.2132	0.6414	0.3156

Table 3. Performance of Bart-base.

Performance			
Model	BLEU \uparrow	Self-BLEU \downarrow	Rouge-L \uparrow
Bart-2 (base) TFIDF			
Top-k			
k = 1	0.1776	0.9000	0.2891
k = 5	0.1890	0.6361	0.2891
k = 10	0.1894	0.5670	0.2966
Bart-2 (base) BM25			
Top-k			
k = 1	0.2295	0.9000	0.3182
k = 5	0.2420	0.5334	0.3265
k = 10	0.2373	0.4715	0.3272

3.5 Spam Filter Passing Rate

We further evaluate whether the emails generated by AI-Phishing can pass NOPAM, a well-known spam filter, and the Gmail spam filter.

First we employ two kinds of emails from sources, (1) Human wrote emails are created from Fraud-dataset (2) AI-Phishing wrote emails from automatic generation, which each chose 30 Fraud-emails and 30 unFraud-emails.

Second, we compare the performance between GPT2 and Bart models on BEC emails using AI-Phishing to compose emails based on trending words.

And next we compare the pass rates of automatically generated emails and emails written by real people on email protection systems. In Table 4, we observe that Normal emails generated by the GPT2 model can successfully pass NOPAM

and Gmail (100% pass rate). It is worth noting that although Bart generates better quality than GPT2, it only has a pass rate of 87% in the Gmail spam filter.

In addition, we also observe the generation of Fraud-type emails. We summarize the results in Table 5. We can see that GPT2 and Human have the same pass rate of 24%, while Bart can achieve a pass rate of 40%. Based on the above two evaluations.

Finally, we compare the BEC email generation pass rate difference between GPT2 and Bart. We observe that GPT2 has a higher pass rate than Bart in Table 6, this mean GPT2 more suitable for the generation of phishing emails in our proposed method, we believe that the emails generated by AI-Phishing have a pass rate close to the emails written by humans.

Table 4. Firewall pass rates (normal).

Firewall pass rates (normal)		
Writer	Firewall	Pass rate
GPT2	NOPAM	100% (30/30)
	Gmail	100% (30/30)
Bart	NOPAM	100% (30/30)
	Gmail	87% (26/30)
Human	NOPAM	100% (30/30)
	Gmail	96% (29/30)

Table 5. Firewall pass rates (fraud/spam).

Firewall pass rates (fraud)		
Writer	Firewall	Pass rate
GPT2	NOPAM	100% (30/30)
	Gmail	24% (08/30)
Bart	NOPAM	100% (30/30)
	Gmail	40% (12/30)
Human	NOPAM	100% (30/30)
	Gmail	24% (08/30)

Case Study. Finally, we design two BEC scenarios (1) financial-related scenarios (2) news current events (COVID-19). The reason for choosing the financial scenario is that most of the BEC attacks usually instruct the recipient to conduct account confirmation and remittances and other issues. As shown in Table 7, the

Table 6. Firewall pass rates (BEC).

Finance related result		
Writer	Firewall	Pass rate
GPT2	NOPAM	100% (30/30)
	Gmail	100% (30/30)
Bart	NOPAM	100% (30/30)
	Gmail	50% (15/30)
COIVD19 related result		
Writer	Firewall	Pass rate
GPT2	NOPAM	100% (30/30)
	Gmail	100% (30/30)
Bart	NOPAM	100% (30/30)
	Gmail	80% (24/30)

Table 7. Case study-finance.

Model : GPT2-Medium
Keywords : 'bank','account','reply','number','credit'
Subject : How was that payment? reply to me soon
Outputs : How did you pay for the **credit** card and how much is in there. I have a **number** of questions about this as well but can't find any information on this issue or even offering assistance to help with it.

Model : GPT2-Medium
Keywords : 'read', 'attached', 'account','bank'
Subject : I will get back to you in the am
Outputs : I will get back to you in the am **bank**, but I don't think I'm going to get a **credit** card. I'm not sure if you are going to be able to get one or two of them.

Model : Bart-large
Keywords : 'read', 'attached', 'account','bank'
Subject : How about was that payment? reply to me soon
Outputs : i am **attached** to this **account** because it is very important for the **bank** to be able to process this money out of a dormant **account**. this **account** is dormant and no one has operated on it again we are still working with the **bank** as we do not know how it works, i will call you back in touch with details thank you best regards

Model : Bart-large
Keywords : 'officer', 'attached', 'account','bank','credit'
Subject : I will get back to you in the am **Outputs :** i have been **attached** to this **account** for years but was never able to do more than a small amount of work on it but now i want to make sure you know what is going on at the end of the day let me start by introducing myself i am an external officer of the **bank** where i work as a **credit** officer. my department has been dealing with various dormant **accounts** ,some are dormant this is due to my inability to reach any of those **accounts** that are directly related to the **account**. however i will like to hear from you soon best regards sending out this message to you

content of the email generated by GPT2 seems to ask the recipient for personal information for the follow-up social engineering process. In comparison, Bart’s generated result is richer(by describing a story and contextual questions).

We think this was because Bart learned the tricks of the original scam email style, using various stories to gain trust. In addition, we also evaluate the generation effect of incorporating news events. As shown in Table 8, we observed that Bart’s report information on COVID19 issues is closer to the news than GPT2. Through the above case study, we can observe that the emails gener-

Table 8. Case study: COVID-19 related news

Model : GPT2-Medium
Keywords :
'covid19', 'case', 'hospital', 'with', 'CDC'
Subject : please, reply to me soon
Outputs : Please respond to my now a few days ago about the hospitalization of his wife. She is recovering from **COVID-19** and has been in ICU for weeks. I am writing you this email because we are both desperate and want help and believe that it's only fair that someone who can assist us should be able do so.

Model : GPT2-Medium
Keywords :
'covid19', 'case', 'hospital', 'with', 'CDC'
Subject : please, reply to me soon
Outputs : Please respond to my now very urgent email, I have not heard from you since last week and am still trying the best that i can **with** no luck so this is going forward as a request for assistance in regards of your **hospitalization** or any additional medical conditions we may deem necessary to treat. Please do let us know if there are other ways which will be more convenient than contacting someone through an alternative means.

Model : Bart-large
Keywords :
'covid19', 'case', 'hospital', 'with', 'CDC'
Subject : please, reply to me soon
Outputs : i am writing you this letter in good faith, i have been diagnosed with **covid19**, a cancer that has defied all forms of medical treatment and hospitalization. i have not particularly lived my life so well but i was fortunate enough to have the opportunity to work **with** the **CDC** where i started my career i now live in cincinnati, i got your contact while searching for a reliable person who can help me find out what is going on at the **hospital**. i want to hear from you soon, thank you best regards.

Model : Bart-large
Keywords :
'covid19', 'case', 'hospital', 'with', 'CDC'
Subject : please, reply to me soon
Outputs : We are keenly interested in setting up a new **hospitalization** program in the near future. We will work **with** our **CDC** to develop a guidelines for the public health lab and are patient-centered care industry to better manage the associated costs.

ated by GPT2 are close to general mail for work correspondence. While Bart's generated quality is good, it is more likely to be detected by mail filters.

4 Conclusion

In this paper, we present *AI-Phishing*, a novel Phishing SET curation system based on natural language generation (NLG) techniques. The core of AI-Phishing is its text generation module which generates mail content based on given keywords. We also explore the employment of multi-field learning to boost model performance. We believe that the use of text generation techniques to aid in the planning of social engineering exercises will be the next interesting natural language generation application.

References

1. Crestani, F., Lalmas, M., van Rijsbergen, C.J., Campbell, I.: "is this document relevant?... probably?" a survey of probabilistic models in information retrieval. *ACM Comput. Surv. (CSUR)* **30**(4), 528–552 (1998)
2. Das, S.D., Basak, A., Dutta, S.: A heuristic-driven ensemble framework for COVID-19 fake news detection. *arXiv preprint arXiv:2101.03545* (2021)
3. Lewis, M., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019)

4. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp. 74–81 (2004)
5. Lin, T., et al.: Susceptibility to spear-phishing emails: effects of internet user demographics and email content. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **26**(5), 1–28 (2019)
6. Misra, R.: News category dataset (2018). <https://doi.org/10.13140/RG.2.2.20331.18729>
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
8. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9 (2019)
9. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. *Nist Spec. Publ. Sp* **109**, 109 (1995)
10. Salloum, S., Gaber, T., Vadera, S., Shaalan, K.: Phishing email detection using natural language processing techniques: a literature survey. *Procedia Comput. Sci.* **189**, 19–28 (2021)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**(5), 513–523 (1988)
12. Verma, A.: Fraud email dataset. kaggle (2018). <https://www.kaggle.com/llabhishekl/fraud-email-dataset>
13. Wolf, T., et al.: Transformers: state-of-the-art natural language processing pp. 38–45 (2020)
14. Zellers, R., et al.: Defending against neural fake news. arXiv preprint [arXiv:1905.12616](https://arxiv.org/abs/1905.12616) (2019)
15. Zhu, Y., et al.: A benchmarking platform for text generation models. arxiv 2018. arXiv preprint [arXiv:1802.01886](https://arxiv.org/abs/1802.01886) (2018)



Stock Price Trend Prediction Using LSTM and Sentiment Analysis on News Headlines

Jung-Bin Li¹, Szu-Yin Lin², Fang-Yie Leu^{3,4}(✉), and Yen-Chu Chu¹

¹ Department of Statistics and Information Science, Fu Jen Catholic University, New Taipei City, Taiwan
071635@mail.fju.edu.tw

² Department of Computer Science and Information Engineering, National Ilan University, Yilan, Taiwan

³ Department of Computer Science, Tunghai University, Taichung, Taiwan
leufy@thu.edu.tw

⁴ Emergency Response Management Center, Ming Chung University, Taoyuan, Taiwan

Abstract. To simulate the trading behavior of investors in the stock market, this study adopts parameters including technical, fundamental, and chip to build a LSTM model, and also observes the ability of news sentiment to predict stock prices. Influential stocks such as TSMC, Fulgent Sun, and HTC are chosen as the target of our experiment. Four common natural language processing packages are used to label news sentiment. Then the combined sentiment labels along with the LSTM model are used for backtesting. The results of the study found that FinBERT's ability to predict the price trend outperforms other methods, with an accuracy of 41.6%. In addition, combining news sentiment labels with the LSTM model generally leads to better outcome than using either the news label or the LSTM model alone. However, in certain extreme cases, traditional technical indicators or even buy-and-hold strategy have better performances.

1 Introduction

Due to the impact of economic and pandemic factors in recent years, coupled with the background factors such as the opening of the Taiwan stock market for intraday fractional trading and the lowering of trading thresholds, more people have been attracted. These novice investors tend to seek simple and efficient ways to make investment decisions. Statistics revealed from the Taiwan Stock Exchange show that by the end of 2021, the total number of accounts opened has exceeded 12 million, and the number of newly opened accounts has reached 770,000, indicating that the general public's willingness to invest in the stock market has increased considerably.

In the past, some studies have used linear regression [1] or Support Vector Machine [2], k-nearest [3] and other models to predict stock prices, and the random walk hypothesis [4], and some also have tried the ARIMA (Autoregressive Integrated Moving Average model) [5]. RNN (Recurrent Neural Network) is one of the common methods in natural language processing. Since stock prices and news posts are convincingly related data, some studies have tried to use LSTM (Long Short-Term Memory) to predict stock

prices [6]. Similar studies in the past mostly focused on English reports and tended to use technical parameters. This research hopes to use more comprehensive information to improve the accuracy of the LSTM model for stock price prediction to assist investors.

2 Related Works

2.1 Long Short-Term Memory (LSTM)

Hochreiter and Schmidhuber proposed LSTM in 1997 [7]. The core concept is that the cell state trades off the incoming information from the hidden layer, so that the problem of dependence between the weights can be solved. Many derivative models of LSTM have been proposed. Graves and Schmidhuber raised Bi-LSTM (Bidirectional LSTM) to analyze phonemes [8]. Greff, Srivastava, Koutnik, Steunebrink, and Schmidhuber tried to change the structure and various parameters of LSTM, and the results showed that the change of its structure could not significantly improve the learning ability of the original model, and the parameters used in fine-tuning were closely related to the training data [9]. Graves, Jaitly, and Mohamed used Bi-LSTM for speech recognition [10], and Shi, Chen, Wang, Yeung, Wong, and Woo applied LSTM to rainfall prediction [11]. Liang, Shen, Feng, Lin, and Yan contributed to the improvement of LSTM in image processing [12].

There are studies applying deep learning to stock price prediction. Liu, Liao, and Ding believe that in the stock price prediction task, the accuracy of multi-layer LSTM will be higher than that of single-layer LSTM, but the cost will also increase [13]. Ojo, Owolawi, Mphahlele, and Adisa used LSTM to predict stock prices and suggested adding factors such as international news and policies as directions for improvement [14]. Bathla [15] compares the stock price prediction ability between LSTM and SVM models, and the results show that LSTM outperforms SVM. Guo added news sentiment labels as input parameters to the LSTM model, and its prediction performance was significantly improved compared to purely using stock daily trading information as input [16]. Kavinnilaa, Hemalatha, Jacob, and Dhanalakshmi use LSTM to predict stock prices, and point out that stock prices are affected by news in the market, and propose that significant information events can be set as parameters to improve model accuracy [17].

2.2 Natural Language Processing (NLP)

Dervlin, Chang, Lee, and Toutanova proposed BERT (Bidirectional Encoder Representations from Transformers) in 2018, which mainly uses the Encoder part of Transformer to generate language models. Since the training of BERT requires a lot of computing resources, most studies use pre-trained models that have been trained for feature extraction as input parameters for fine-tuning training [18]. Araci screened keywords in the financial field and established a TRC2-financial dataset, which contains more than 29,000,000 words and nearly 400,000 sentences, making FinBERT a pre-trained BERT model in the financial field [19]. Therefore, this study applies FinBERT and other NLP methods with the attention to reach good results.

3 Proposed Model and Mechanisms

3.1 Model Framework

In this study, the parameters were first imported from the relevant databases, and the data was preprocessed. Then we use the LSTM model to train and compare the differences among parameter combinations. Our LSTM model takes the daily closing price of a stock as the dependent variable, and the independent variables are retrieved from different perspectives. In addition, we observe whether the sentiment of news headlines is helpful for stock price prediction, and then use the output of the model as the basis for stock operation to analyze and compare the backtest results. Model framework is illustrated in Fig. 1.

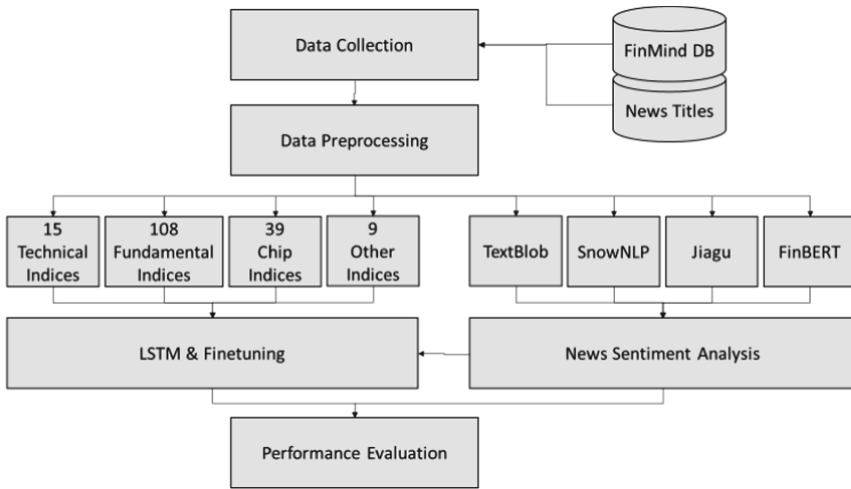


Fig.1. Model framework.

3.2 Data Source

This study uses data from the FinMind database to obtain 171 variables from technical, fundamental, and chip perspectives; news sources are obtained from 23 online news databases. We use these variables to form the reference factors for investment decisions.

Technical. Using the stock price daily transaction information (8), individual stock PER and PBR data (3), the day’s offset transaction target and trading volume value (3), and the weighted stock price index, a total of 15 variables are used. The numbers in brackets represent the number of variables adopted. The daily transaction information table of stock prices includes transaction volume, transaction amount, opening price, highest price, lowest price, closing price, the difference from the previous day’s closing price, and the number of transactions. The PER and PBR data of individual stocks include cash

yield, price-to-earnings ratio, and price-to-book value ratio. The target and volume value of the offset transaction on the day include the transaction volume, the buying amount, and the selling amount.

Fundamental. A total of 108 variables are obtained from the consolidated income statement (25), balance sheet (44), cash flow statement (33), dividend policy and ex-rights and ex-dividend results (4), and monthly income statement (2).

Chip. A total of 39 variables were obtained from the margin financing and securities lending table (11), legal person trading (5), corporate trading in the Taiwan market as a whole (6), foreign shareholding (2), and shareholders' shareholding classification table (15).

News. Stock-related Mandarin news headlines from 23 sources including Anue, EBC, ETtoday, HiNet News Community, INSIDE, MoneyDJ, NOWnews, UDN, YahooNews, ChinaTimes, BusinessToday, TaiwanNews, AppleDaily, BusinessWeekly, Commonwealth, CommercialTimes, NewTalk, MoneyWeekly, TechNews, ManagerToday, LibertyTimes, Sina, and GlobalViews are collected.

Others. Nine variables including the exchange rate of Taiwan dollar against the US dollar, the US Federal Reserve interest rate, the US one-month Treasury bond yield, the price of gold, the price of crude oil, the CNN Fear and Greed Index, the US stock TSM ADR, the Dow Jones Index, and the Philadelphia Semiconductor Index are adopted.

3.3 Data Processing

Preprocessing. Because news reports are usually published outside the stock trading day, we fill in its 30-day moving average if the stock information does not exist during the unopened period. Quarterly data will be repeatedly filled downwards until the next season's release date.

News Sentiment Labels. This research uses TSMC (2330), which is the largest market capitalization in Taiwan stock market, as the main object. Benefiting from TSMC's trading volume, its related news is ample. The research retrieves news from July 28, 2020 to May 1, 2022, with a total of 10,381 records. Cleaning works including removing duplicate titles, stock codes, news webpage links, data sources and the data source part in the title are done.

This study applies different methods to label news headlines with sentiment. The first method is to use TextBlob to get news headline sentiment scores. TextBlob is a text processing tool based on two python packages, NLTK and Pattern. The TextBlob output format is as follows:

Sentiment (polarity = 0.3916, subjectivity = 0.4357)

The polarity score ranges from -1 to 1 . The closer to -1 , the more pessimistic the sentiment is; the closer to 1 , the more optimistic the sentiment. Subjectivity ranges from 0 to 1 , the closer it is to 0 , the more objective it is.

The second is SnowNLP, which is similar to TextBlob. The training of sentiment analysis tasks uses the Bayesian model and its self-equipped dictionary. Its score represents the probability of being a positive sentiment, ranging from 0 to 1. The closer the value is to 1, the more positive it is.

The third is Jiagu. It also supports the function of customizing the new dictionary, but is based on BiLSTM training. After Jiagu word segmentation, the idea of bag of words is used, and the NTUSD emotional vocabulary dictionary from the Taiwan Natural Language Processing Laboratory is used for classification. The negative words are given a score of -1, and the positive words are given a score of 1. For neutral words, a score of 0 is given.

The fourth is to use FinBERT to get news headline sentiment scores. FinBERT is a pre-trained model for analyzing financial text sentiment. Its output format is as follows: { 'label': 'Negative', 'score': 0.9966173768043518 }

The label has three possible sentiments: negative, positive, and neutral. Negative words are classified to negative and given a score of -1; positive words are classified to positive and given a score of 1; and neutral words are classified to neutral. With score 0. The score represents the intensity of sentiment. This study adopts the label column only.

3.4 Data Analysis

In this study, TSMC (2330) is chosen to evaluate our model architecture. In order to observe the model performance, we additionally take Fulgent Sun-KY (9802) and HTC (2498) as supplement for the analysis. This research collected a total of 603 data tuples from July 28, 2020 to May 1, 2022. The input fields cover technical, fundamental, chip, news, and other variables, with a total of 171 variables. The data is divided into training data set and verification data set with the ratio 9 to 1.

Subsequently, we use the TSMC news data to compare the prediction ability of each news sentiment labeling method, and then use the most effective labeling method to train the LSTM model and back-test for comparison.

4 Model Performances and Findings

4.1 LSTM Outcome

This study trains and compares models using different input and output lengths, including the following combinations: 1-to-1, 1-to-5, 5-to-1, 5-to-5, 30-to-1, and 30-to-5. One-to-one features the model of using data of one day in the past to predict the price trend of one day in the future. To simplify the description, subsequent references to the model will denote the three-layer 1-to-1 LSTM as LSTM3 (1, 1).

All LSTMs use Adam (Adaptive Moment Estimation) as the Optimizer, and the dropout ratio in the architecture is set to 0.2. We add Early Stopping to prevent overfitting, and training will be stopped when the loss has not been improved after 10 epochs. The comparison between the models is based on RMSE, and the results are shown in Table 1.

Table 1. Comparison of LSTM models.

Model	RMSE	Model	RMSE
LSTM1 (1, 1)	33.2822	LSTM1 (5, 5)	64.3136
LSTM3 (1, 1)	20.2428	LSTM3 (5, 5)	50.9448
LSTM5 (1, 1)	20.3868	LSTM5 (5, 5)	52.1754
LSTM1 (1, 5)	131.4313	LSTM1 (30, 1)	35.6099
LSTM3 (1, 5)	32.7369	LSTM3 (30, 1)	27.8210
LSTM5 (1, 5)	40.3517	LSTM5 (30, 1)	28.6284
LSTM1 (5, 1)	42.8656	LSTM1 (30, 5)	49.8965
LSTM3 (5, 1)	48.6941	LSTM3 (30, 5)	38.8199
LSTM5 (5, 1)	51.9153	LSTM5 (30, 5)	34.3075

In this part of the experiment, the following findings were observed:

1. Increasing the number of layers of the LSTM model did not always improve the performance of the mode. There is little difference between the three-layer and five-layer LSTM. Considering the computational efficiency, subsequent experiments will be carried out with three-layer LSTM.
2. The multi-to-1 models are not as good as using the 1-to-1 model. Among our experimental models, the LSTM3 (1, 1) model has the best predictive ability (RMSE = 20.2428).
3. The results of multi-to-multi and less-to-more models are in average worse than other combinations, and the maximum RMSE is from LSTM1 (1, 5).

4.2 News Sentiment

In the news sentiment analysis module, this research uses four sentiment labels, namely FinBERT, Jiagu, TextBlob, and SnowNLP, to classify a total of 10,385 TSMC news headlines from July 28, 2020 to May 1, 2022. In order to understand the distribution of the data, the results of each label are further classified into interval scales. FinBERT uses its output directly. Jiagu is set neutral when score 0, positive when the score is greater than 0, and negative when less than 0. TextBlob is set neutral when score equals to 0, positive when greater than 0, and negative when less than 0. SnowNLP is neutral when scores are in the range (0.4, 0.6), positive when greater than 0.6, and negative when less than 0.4. The distribution is shown in Table 2.

All labeling methods except SnowNLP have more neutral labels. SnowNLP has more negative labels, and it may be attributed to its dictionary which is majorly based on e-commerce. In the first half of the study's training period, TSMC has a general rise trend. Hence the news media reported mostly good news, and the distribution of news labels fits the actual situation.

Table 3 shows the overall accuracy of news labels. FinBERT owns the highest accuracy, and SnowNLP is the lowest. In terms of precision, again FinBERT is slightly

Table 2. Distribution of news labels.

Method	FinBert	Jiagu	TextBlob	SnowNLP
Positive	2555	3394	4170	2085
Neutral	5981	5134	4463	829
Negative	1849	1857	1752	7471

higher than other methods in neutral and negative news classification. However, Jiagu is slightly higher than FinBERT in the matter of positive news. In terms of recall, Jiagu is the highest for positive news, FinBERT for neutral, and SnowNLP for negative. Jiagu may be benefited by its customizable positive dictionary. The number of negative news classified by SnowNLP is significantly higher than that of other methods, which causes its higher recall.

Table 3. Comparison of news sentiment labels.

Method		FinBert	Jiagu	TextBlob	SnowNLP
Accuracy		0.416177	0.379104	0.36564	0.260664
Precision	Positive	0.260281	0.280022	0.220320	0.243073
	Neutral	0.534024	0.509349	0.532153	0.496984
	Negative	0.252838	0.236299	0.245803	0.229736
Recall	Positive	0.192800	0.208000	0.154400	0.012630
	Neutral	0.590825	0.483721	0.439327	0.093049
	Negative	0.260589	0.323518	0.413473	0.479391

4.3 Model Backtest

This study collects the price data of TSMC from March 3, 2022 to May 1, 2022. The initial holding capital is set at 10 million NTD. Other assumptions include the securities transaction tax as 0.3% of the transaction amount, and the brokerage fee is 0.1425% of the transaction amount. We use the 1-to-1 3-layer LSTM model. When the predicted value is greater than the actual stock price, it is seen as a buy signal. When the predicted value is less than the actual stock price, it is set as a sell signal. When the predicted value is equal to the actual stock price, we hold without trading.

Then the FinBERT news sentiment label is used as an operation reference. We buy when the sentiment label is positive; we sell when it is negative; and we hold when it is neutral. Additionally, we make decisions based on the FinBERT news sentiment label combined with the result of the LSTM model. It triggers a buy signal when the sentiment label is positive and the LSTM prediction result is greater than the actual stock price.

Sell signals ring when the sentiment label is negative and the LSTM prediction result is lower than the actual stock price.

The backtest results of the above strategies are summarized in Table 4. All models end up with negative earnings. The reason is that TSMC's price trend is downward during the testing period, indicating that the overall trend of the investment target has a direct impact on the final profit, and its timing of buying is equally important. However, for the three models of LSTM, news sentiment label, and LSTM plus sentiment label, the LSTM plus sentiment label model is better than the others.

Table 4. Backtest results of TSMC.

Model	LSTM	News label	LSTM + News label	Buy & hold
Average profit (NTD)	-81754	-76106	-44214	-22079
Max loss (NTD)	724128	283924	258855	62693
Average profit rate (%)	-0.82	-0.76	-0.44	-0.22
Net profit margin (%)	-4.36	-2.03	-1.74	-0.46
Internal rate of return (%)	-23.75	-11.73	-10.13	-2.77
Sharpe index	-1.34	-2.66	-2.24	-1.81

To further validate the performance of our model, Fulgent Sun (9802), which had an upward stock price trend, and HTC (2498), which fluctuated frequently, were used as comparisons. The results are shown in Table 5. For the backtest of 9802, the performance ranking is news, LSTM plus news, and LSTM. The result is different from that of TSMC. HTC's backtest results are similar to those of TSMC, and LSTM and news sentiment labels cannot capture the starting point of the stock price. But what is certain from this

Table 5. Back test results of 9802 and 2498.

StockID	9802			2498		
	LSTM	News label	LSTM + News label	LSTM	News label	LSTM + News label
Average profit (NTD)	78912	260858	154183	-424085	-39075	-23263
Max loss (NTD)	68120	67430	54173	157342	167224	97219
Average profit rate (%)	0.79	2.61	1.54	-0.42	-0.39	-0.23
Net profit margin (%)	6.44	11.82	8.2	-1.43	-1.35	-0.8
Internal rate of return (%)	37.22	76.18	49.11	-7.14	-6.75	-4.05
Sharpe index	4.87	6.18	5.69	-2.9	-2.42	-2.5

performance comparison is that adding news sentiment labels does have a positive effect on stock price prediction. This finding stands in the backtest of all three stocks of the study.

5 Conclusion

The first part of this study tried to use LSTM models with different layers to predict the stock price of TSMC (2330), and found that the 1-to-1 3-layer model outperformed other combinations, and the 1-to-1 5-layer model had a slightly higher RMSE than 3-layer model. The 30-to-1 model, on the other hand, is not as predictive as the 1-to-1 model. It is speculated that some parameters generate noise on the prediction results during learning, which has an impact as the sampling time lengthens. In the comparison of different news sentiment labels, including the prediction of stock price fluctuations, the accuracy rate of FinBERT is the highest, which is 41.6%. This is not good enough in our experiment, so we have to give investors some reminder that it is not a reliable reference to make investment decisions solely based on news.

For 2330, all three models in the backtesting stage ended up with negative profits, because its stock price trended down significantly during the testing period. Compared with the LSTM model, the LSTM plus news labels model has more restrictions on buying and selling conditions, and it is less likely to trigger buying and selling behavior and reduce losses. The LSTM and LSTM plus news labels models continued to buy during the period, and the profit situation was similar to the buy-and-hold method. The unrealized loss of the news model is smaller than that of LSTM and LSTM plus news labels, but the realized profit and loss of LSTM and LSTM plus news labels is higher than that of news. Regarding to the average profit result, LSTM plus news labels is better than LSTM and news strategy.

Compared with 2330, the results of Fulgent Sun (9802) and HTC (2498) vary in model performance ranking. However, it can be observed that the addition of the news sentiment label model does help profit. This outcome inspires us to further extend and fine-tune our model.

References

1. Bhuriya, D., Kaushal, G., Sharma, A., and Singh, U.: Stock market predication using a linear regression. In: 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 510–513 (2017)
2. Lai, L.K.C., and Liu, J.N.K.: Stock forecasting using support vector machine. In: 2010 International Conference on Machine Learning and Cybernetics, pp. 1607–1614 (2010)
3. Taunk, K., De, S., Verma, S., and Swetapadma, A.: A brief review of nearest neighbor algorithm for learning and classification. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255–1260 (2019)
4. Fama, E.F.: Random walks in stock market prices. *Financ. Anal. J.* **21**(5), 55–59 (1965)
5. Ariyo, A.A., Adewumi, A.O., and Ayo, C.K.: Stock price prediction using the ARIMA model. In: 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, pp. 106–112 (2014)

6. Chen, K., Zhou, Y., and Dai, F.: A LSTM-based method for stock returns prediction: a case study of China stock market. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 2823–2824 (2015)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), Montreal, QC, Canada, Vol. 4, pp. 2047–2052 (2005)
9. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(10), 2222–2232 (2017)
10. Graves, A., Jaitly, N., and Mohamed, A.: Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278 (2013)
11. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K. and Woo, W.C.: Convolutional LSTM Network: a machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015), vol. 1, 802–810 (2015)
12. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph LSTM. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 125–143. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_8
13. Liu, S., Liao, G. and Ding, Y.: Stock transaction prediction modeling and analysis based on LSTM. In: 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 2787–2790 (2018)
14. Ojo, S.O., Owolawi, P.A., Mphahlele, M., Adisa, J.A.: Stock market behaviour prediction using stacked LSTM networks. 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), pp. 1–5 (2019)
15. Bathla, G.: stock price prediction using LSTM and SVR. In: 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 211–214 (2020)
16. Guo, Y.: Stock price prediction based on LSTM neural network: the effectiveness of news sentiment analysis. In: 2020 2nd International. Conference on Economic Management and Model Engineering (ICEMME), pp. 1018–1024 (2020)
17. Kavinnilaa, J., Hemalatha, E., Jacob, M.S., Dhanalakshmi, R.: Stock price prediction based on LSTM deep learning model. In: 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–4 (2021)
18. Dervlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
19. Araci, D.: FinBERT: financial sentiment analysis with pre-trained language models. In: The International Conference on Learning Representations (ICLR) arXiv: 1908.10063v1 cs.CL (2019)



Efficient Weighted and Balanced Resource Allocation for High-Performance Render Farms

Lung-Pin Chen¹, Fang-Yie Leu^{1,2(✉)}, Chia-Chen Kuo³, Tzu-Ching Lin¹,
and Ming-Jen Wang³

¹ Department of Computer Science, Tunghai University, Taichung City, Taiwan
{lbchen, leufy, g09350021}@thu.edu.tw

² Emergency Response Management Center, Ming Chung University, Taoyuan City, Taiwan

³ National Center for High-performance Computing, Hsinchu, Taiwan
{c00kcc00, reddy}@narlabs.org.tw

Abstract. To speed up the rendering tasks that are both compute-intensive and I/O-intensive, the render farm is constructed to execute rendering jobs in parallel. The weighted scheduling approach is widely used in the render farm industry such as AWS to address the issue of scalability and robustness. In this paper, we extend the weighted scheduling to the adaptive environment with the changing number of users, jobs, and compute nodes. Since modern Cloud render farms are designed to accommodate various types and scales of users and jobs, our work in this paper can preserve the advantages of the weighted scheduling approach while extending to the Cloud environment.

1 Introduction

The continuous growth of global internet users and evolving of modern communication technologies are driving the demand for digital media applications such as movies, games, and VR (virtual reality). Computer rendering is a process of constructing images from a collection of models and scene files [3]. An animation/video is comprised of *frames i.e.* a single image in a sequence of pictures. Rendering a visually realistic frame can take several hours or even several days. A typical animation/video has 24 or 30 frames per second. Therefore, rendering and constructing image frames is a compute-intensive process. Rendering is also IO-intensive since it needs to read the large scene data files from the storage over the high-bandwidth network and write the result back after rendering.

Since the computation of rendering a frame usually does not depend on another frame, rendering jobs are highly parallelizable. A *render farm* [1, 4, 6] refers to a high-performance computer cluster that is constructed intended to speed up rendering by distributing the tasks to the cluster of computers.

As a render farm is designed to accommodate diverse rendering applications, managing computational resources and assigning jobs to the computers have become major challenges for a render farm [2, 5, 6]. Many render farms use the pay-on-demand system [1]. In such a system, the priority of user jobs is positively related to user payment. That is, the more the user pays, the higher the priority of its jobs.

The advantage of using priority values to determine the resource allocation in a render farm is simplicity and robustness [1]. However, there are some drawbacks to the traditional priority-based method. First, the priority of a user is calculated mainly based on customer payment. The system environment parameters are not involved in the step of calculating priorities, making it not adaptive to the render farm of different scales. Moreover, the arrival rate of user jobs is usually not uniform. Mostly, a user submits jobs from time to time that follows a Poisson distribution. In such a case, a high-priority user may require compensation for idle time. But the traditional priority-based method [1] only allocates resources by considering the priority value evaluated at the current time.

This paper addresses the above issues of the weighted scheduling used in a render farm. We developed a new weighted scheduling algorithm by incorporating more system environment-related parameters into the weight evaluation formula. The new approach enables adaptive resource allocation for the render farm of different scales and also solves the starvation problem.

The rest of this paper is organized as follows. In Sect. 2, we will introduce the backgrounds. Section 3 will introduce our new weighted scheduling algorithm. Section 4 is the experiments section. Finally, concluding remarks and future work are given in Sect. 5.

2 Backgrounds

This section will explain the architecture of the render farm as well as the weighted scheduling algorithm based on user priorities.

2.1 Render Farm

Rendering is a computer-generated image technology in which the computer calculates the color of each pixel based on materials, textures, and lighting that are described in the scene file. Computer image generation can be divided into two types: pre-rendering and real-time rendering [4].

Pre-rendering is often used in visual effects and 3D animations. Each picture is pre-computed, and then the picture is played at a certain rate. This method requires a lot of computing power for simulation, but it can generate realistic images that are difficult to be distinguished by human eyes. Real-time rendering is often used in games and focuses on user interaction. To generate images in a very short time, some visual effects such as lighting and shading will be sacrificed to meet the requirements of real-time computing.

AWS Deadline [1] is a commercial render farm manager for Windows, Linux, and mac OS. As depicted in Fig. 1 (Left), a Deadline system comprises four components: Database, Repository, workstation (*i.e.* servers), and render nodes (*i.e.* workers). The Database stores the jobs, settings, and worker configurations. The Repository stores the plugins, scripts, logs, and any auxiliary files that are associated with the jobs. The render nodes, *i.e.* workers, connect to Database and Repository to fetch and run the rendering jobs. Finally, the workstations are servers performing dedicated functionalities such as monitoring and controlling the rendering processes run on workers. Some other workstations perform dedicated functionalities including auto-configuration,

power management, worker throttling, statistics gathering, license management, and so on.

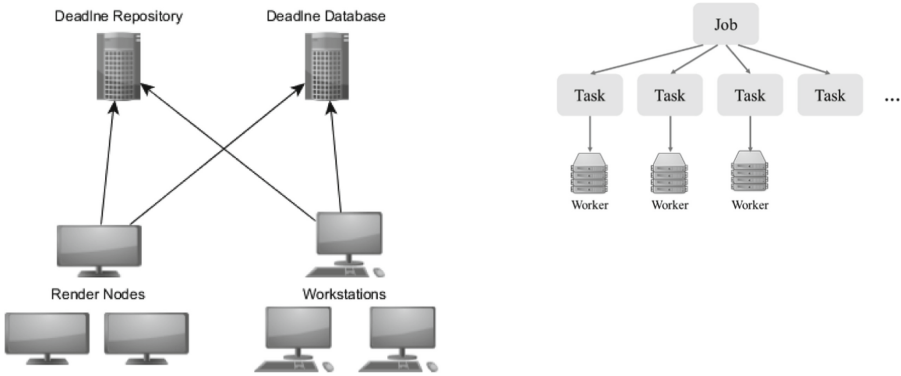


Fig. 1. (Left) The architecture of Deadline render farm management system. (Right) Jobs and its constituent tasks.

Figure 1 (Right) shows the job breakdown in the Deadline render farm. A *task* represents the computation of rendering one or more frames. A *job* consists of multiple rendering tasks. These tasks can be assigned to multiple nodes to compute in parallel. A job is completed when all its constituent tasks are completed.

2.2 Weighted Scheduling Based on Priorities

In a render farm, the *weighted scheduling algorithm* determines a job weight by taking into account the four types of parameters: user priority, submission time, the number of rendering tasks (belonging to the same job), and the number of job errors. As described in [1], the submission time will be the tie-breaker if two jobs have the same weight.

For each job *j*, its weight is calculated as follows [1]:

$$\begin{aligned}
 \text{weight}(\text{job } j) = & (j.\text{Priority} * PW) + \\
 & (j.\text{Errors} * EW) + \\
 & ((NOW - j.\text{SubmissionTimeSeconds}) * SW) + \\
 & ((j.\text{RenderingTasks} - RB) * RW)
 \end{aligned} \tag{1}$$

In the above formula, *j.Priority* refers to the job priority which is determined by the price paid by the user who owns job *j*. Also, *NOW* refers to the current system time and *j.SubmissionTimeSeconds* refers to the submission time of job *j*, both are measured in seconds. The *j.RenderingTasks* refers to the number of tasks in job *j* (see Fig. 1 (Right)) that are currently under execution.

When calculating the total job weight, we can control how much each item contributes to the total value by using the following weight parameters:

- **PW** = the priority weight
- **SW** = submission time weight
- **RW** = rendering task weight

The rest parameters **RB** and **EW** are not directly related to our extended method and are set to zero hereinafter in this paper.

After the weights of all jobs are calculated, the scheduler chooses a job with the largest weight to assign to a worker for execution. Note that the weight parameters can be positive or negative. For example, we usually set **PW** and **SW** to positive. This means that users who wait longer have higher priority. In the meantime, we usually set **RW** to *negative* which means that a job that gains more resources will have lower priority so that other jobs can obtain resources with a high probability. The system will de-prioritize the jobs that have already received more resources to improve the chances of other jobs getting service.

3 New Weighted Scheduling Algorithm

In this section, we first investigate the weakness of previous weighted scheduling. We then propose a new algorithm to address these problems.

3.1 Limitations of Traditional Weighted Scheduling Method

When a worker is idle, it sends a task request to the master server. The master then selects and assigns a job in queue to this worker. The jobs are selected according to the weighted scheduling formula *i.e.* Equation (1). To be more specific, let us examine the example shown in Fig. 2.

In the example of Fig. 2, there are user A (owns job A1 with $A1.Priority = 60$), user B (owns job B1 with $B1.Priority = 40$), and 40 workers in the render farm. The weight parameters in the scheduling formula (1) are set to $PW = 1$, $EW = 0$, $SW = 0$, $RW = -1$. Note that all the parameters are positive while only RW is negative. In the table of Fig. 2, each row represents the task assignment for one scheduling request. The left column represents the id of the worker issuing the task request. The middle column represents the current weight of job A1 and B1. The right column represents the current resources (*i.e.* workers) gained by job A1 and B1.

As shown in the table of Fig. 2, initially, the weights of A1 and B1 were equal to 60 and 40 respectively. Therefore, A1 has a higher weight and will get served for the next task request. Each time a new task in job A1 gets executed, the counter $A1.RenderingTasks$ is incremented by one. Since $RW = -1$ and $RB = 0$, the counter increases by one causing the weight item $(A1.RenderingTasks - RB) * RW$ to decrease by one. Therefore, as shown in Fig. 2, during the scheduling steps 1st to 19th, job A1 obtains 20 workers in total and its weight dropped from 60 to 41. Then,

worker	weight		RenderingTasks	
	A1	B1	A1	B1
	60	40	0	0
1 st worker	59	40	1	0
2 nd worker	58	40	2	0
...				
19 th worker	41	40	19	0
20th worker	40	40	20	0
21st worker	40	39	20	1
22 nd worker	39	39	21	1
23 rd worker	39	38	21	2
24 th worker	38	38	22	2
...				
40 th worker	30	30	30	10

Fig. 2. Task assignment of job A1, B1, and 40 workers.

in scheduling step 20th, both job A1 and B1 have the same weight i.e. 40. Starting from step 20th, job A1 and B1 take turns to allocate resources.

According to our practical experience, the above-mentioned weighted scheduling inherently incurs several limitations. First, the weights are calculated *based on jobs* rather than *users*. If one high-priority user submits two or more jobs, all of these jobs will inherit the high priority since the parameter `job.RenderingTasks` is calculated for each job individually. In other words, a single user can issue multiple jobs to the system queue to gather more resources. Second, the weights are calculated by using the parameters that are evaluated in integer values rather than relative ratios. This makes the resource assignment induced by formula (1) cannot be adapted to the computer cluster of different scales. The final problem of the previous weighted scheduling is the average rate of long-term resource allocation. Usually, the arrival rate of user jobs is not uniform. Mostly, a user submits jobs from time to time that follows a Poisson distribution. In such a case, a high-priority user may require a compensation idle period. However, the previous priority-based method [1] only allocates resources by considering the priority value that is evaluated at the current time.

worker	Weight			RenderingTasks			UserTotalWorker	
	A1	A2	B1	A1	A2	B1	userA	userB
	60	60	40	0	0	0	0	0
1 st worker	59	60	40	1	0	0	1	0
2 nd worker	59	59	40	1	1	0	2	0
...								
40th worker	40	40	40	20	20	0	40	0

Fig. 3. Task assignment of job A1, A2, B1, and 40 workers.

Figure 3 illustrates the above problems of the weighted scheduling method. Similar to the example in Fig. 2, there are user A, user B (with `B1.Priority = 40`), and 40 workers in the render farm. The difference is that user A owns two jobs A1 and A2, both with `A1.Priority = A2.Priority = 60`. The workers are assigned to tasks of the jobs according to Eq. (1) as described in Fig. 2. As the scheduling result shown in Fig. 3, both jobs A1 and A2 have high priority than B1 and thus gather all the 40 workers in a monopoly manner, leading to a situation of starvation of user B.

3.2 New Weight Scheduling Algorithm

This subsection addresses the problems mentioned in the previous subsection. We proposed the formula of the extended weighted scheduling as follows:

$$\begin{aligned} \text{weight}(\text{user } u, \text{job } j) = & \\ & (j.\text{RPriority} * \text{PW}) + \\ & (j.\text{Errors} * \text{EW}) + \\ & ((\text{NOW} - j.\text{SubmissionTimeSeconds}) * \text{SW}) + \\ & ((u.\text{RenderingTasks} - \text{RB}) * \text{RW}) \end{aligned} \quad (2)$$

where `j.RPriority` refers to the priority ratio of job `j` that is normalized by the current number of *active users* `M`. A user is considered *active* if it has jobs executed by some workers in the render farm recently *e.g.* within an hour. Specifically, for job `j`:

$$j.\text{RPriority} = M \times p_j / \sum_i p_i \quad (3)$$

where `pj` represents the original priority of job `j`.

Unlike Eq. (1) which uses `j.RenderingTasks`, Eq. (2) uses the other parameter `u.RenderingTasks` which is defined in Eq. (4): for each user `u`:

$$u.\text{RenderingTasks} = \sum_{\text{job } j \text{ owned by user } u} j.\text{RenderingTasks} \quad (4)$$

The parameter `u.RenderingTasks` refers to the total number of workers that are assigned to user `u`.

4 Experiments

In this work, we implement the main components of the render farm by using the OMNet++ simulation toolkit. In this section, we discuss three experiments to demonstrate the effect of our new extended weighted scheduling algorithm, as described in Subjects. 4.1, 4.2, and 4.3.

4.1 Monopoly of Resource Allocation

This experiment is designed to demonstrate the resource monopoly by a single user with high property in the computer cluster. There are 100 workers and four users A, B, C, and D with priority values of 40, 30, 20, and 10, respectively. Users A, B, C, and D have respectively 30, 20, 10, and 10 jobs, and each job contains 100 tasks. We assume that every user adopts the strategy which tries to gather as many resources as possible by submitting multiple jobs. That is, at the beginning time, user A submits all the 30 jobs at once; so as do the other users. The scheduler performs the weighted scheduling with Eq. (1).

The scheduling result of the first experiment is shown in Fig. 4. We can easily observe the resource monopoly as mentioned in Subsect. 3.1 (Fig. 3). As shown in Fig. 4, by launching multiple jobs, the user with the highest priority (*i.e.* user A) will gather all the resources until all its jobs are consumed.

4.2 Starvation Occurred in Weighted Scheduling

This experiment is designed to demonstrate the situation of starvation for the user with the lowest property when using the old scheduling algorithm *i.e.* Eq. (1). There are 50 workers and four users A, B, C, and D with priority values of 40, 30, 20, and 10, respectively. In this experiment, we assume that each user submits only one job at a time. A new job will not be launched until the previous one completes. Each job contains 100 tasks.

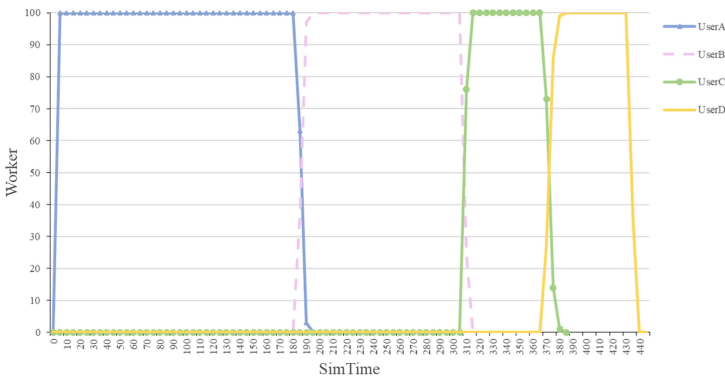


Fig. 4. The scheduling result of the experiment described in Subsect. 4.1. The X-axis represents times (in seconds). The Y-axis represents the number of workers that are allocated for each user.

The scheduling result of the old weighted scheduling with Eq. (1) is shown in Fig. 5. We can observe the situation of starvation for user D. Before 130 s, since there are not enough workers for all four users, user D with the lowest priority cannot gather any worker. Then, after 130 s, since user A starts to finish executing jobs and release the occupied worker resources, the other users B, C, and D can share more resources.

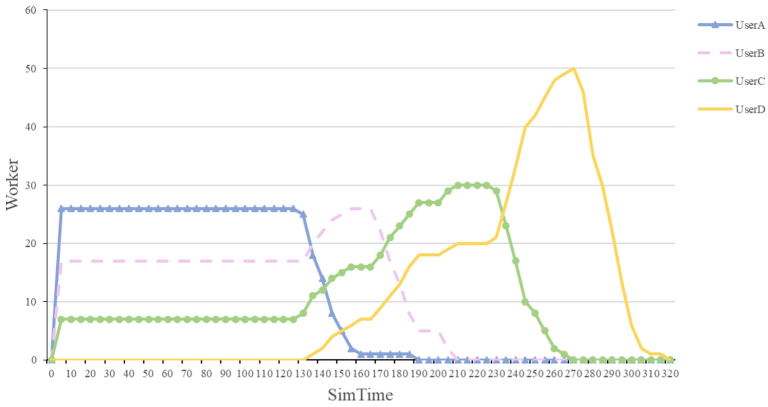


Fig. 5. The scheduling result of the experiment described in Subject. 4.2.

4.3 Improved Extended Weighted Scheduling

This subsection presents the result of the experiment of the new extended weighted scheduling proposed in this paper. The extended scheduling calculates weights by using Eq. (2). In this experiment, there are 50 workers and four users A, B, C, and D with a priority ratio 4:3:2:1.

In this experiment, similar to the experiment in Subject. 4.1, we assume that every user adopts the strategy which trying to gather as many resources as possible by submitting multiple jobs. Thus, similar to the experiment in Subject. 4.1, Users A, B, C, and D respectively launch 30, 20, 10, and 10 jobs at the beginning time. Each job contains 100 tasks.

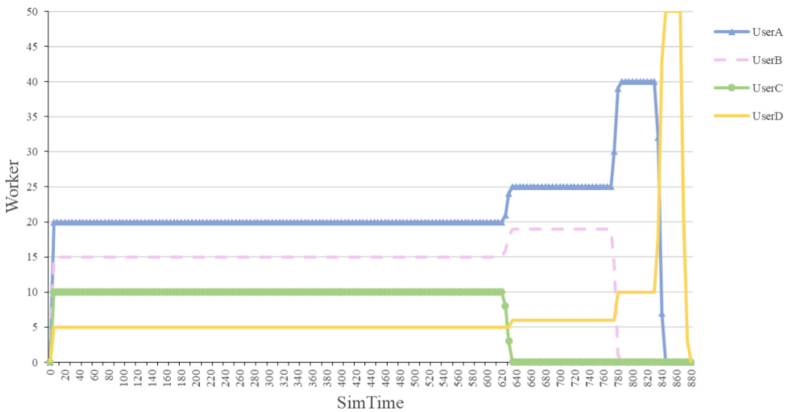


Fig. 6. The scheduling result of the experiment described in Subject. 4.3.

The scheduling result of using the new weighted scheduling with Eq. (2) is shown in Fig. 6. We can observe that users A, B, C, and D share the resources (*i.e.* workers) in a

ratio of 4:3:2:1. Notably, after 600 s, user C completes all its jobs and, in the meanwhile, the other three users share the resources in the ratio normalized by $M = 3$. According to Eqs. (3) and (4), the number of active users is $M = 4$ from 0 to 599 s, while $M = 3$ after 600 s. The result in Fig. 6 demonstrates the adaptability of the new scheduling algorithm to the dynamic changing number of users. It also solves the problem of starvation when there are not enough workers in the system which is appeared in the second experiment in Subsect. 4.2.

5 Conclusion

This paper investigated several problems of the previous weighted scheduling method that is widely adopted in many commercial Cloud render farms. We designed a new extended weighted scheduling by considering the adaptive environment with the changing number of users, jobs, and compute nodes. We implemented the components of the render farm by using the OMNet++ simulation platform. We conducted three experiments in the simulation platform to demonstrate the effect of our new extended weighted scheduling algorithm. The experimental result demonstrates the adaptability of our new algorithm to the dynamic changing number of users. It also solves the problem of starvation when using the old weighted scheduling algorithm.

References

1. AWS Thinkbox: Job scheduling (2020). https://docs.thinkboxsoftware.com/products/deadline/10.1/1_UserManual/manual/job-scheduling.html
2. Schwarzkopf, M., Konwinski, A., Abd-El-Malek, M., Wilkes, J.: Omega: flexible, scalable schedulers for large compute clusters. In: Proceedings of the 8th ACM European Conference on Computer Systems, pp. 351–364 (2013)
3. Savage, T.M., Vogel, K.E.: An Introduction to Digital Multimedia. Jones & Bartlett Publishers, Burlington (2014)
4. Sheharyar, A., Bouhali, O.: A framework for creating a distributed rendering environment on the compute clusters. *Int. J. Adv. Comput. Sci. Appl.* **4**(6) (2013)
5. Chen, L.P., Wu, I.C., Liang, G.Z.: Enhancing parallel game-tree searches by using idle resources of a high performance render farm. In: 2015 Conference on Technologies and Applications of Artificial Intelligence (2015)
6. Yao, J., Pan, Z., Zhang, H.: A distributed render farm system for animation production. In: Natkin, S., Dupire, J. (eds.) ICEC 2009. LNCS, vol. 5709, pp. 264–269. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04052-8_31



A Brake Assisting Function for Railway Vehicles Using Fuzzy Logic: A Comparison Study for Different Fuzzy Inference Types

Mitsuki Tsuneyoshi¹, Makoto Ikeda²(✉), and Leonard Barolli²

¹ Graduate School of Engineering, Fukuoka Institute of Technology,
3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan
mgm21106@bene.fit.ac.jp

² Department of Information and Communication Engineering, Fukuoka Institute
of Technology, 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan
makoto.ikd@acm.org, barolli@fit.ac.jp

Abstract. For trains, cars and other vehicles, the quality of the user comfort level is affected significantly by the driver skill level such as agility, experience and physical condition. In order to assist the drivers, in this paper, we propose a brake assist function for railway vehicles using Fuzzy Logic. We present a comparison study for different Fuzzy Inference types (models). The proposed function provides an intelligent braking for the train drivers considering velocity, current brake level and environment status on the railway. We evaluate four Fuzzy inference types. The simulation results show that Type-2 FLC2 can provide better brake assisting function for train drivers than other models.

Keywords: Brake assist function · Type-2 Fuzzy · Sugeno Fuzzy · Railway vehicle

1 Introduction

Recently, Electronic Control Units (ECUs) installed in different vehicles are transitioning to digital management systems [1] to enhance convenience, safety and compliance with rules on exhausting gas emissions [17]. Also, the number ECUs installed in railway vehicles is growing rapidly. Furthermore, modern cars integrated with adaptive cruise control, collision damage mitigation braking, and lane departure warning systems have been equipped with an Advanced Driver Assistance System (ADAS).

Railway vehicles do not require advanced automatic operating controls like cars. In 1981, unmanned, fully-automated trains began running in Japan, following a guide track on a dedicated track. Because they travel on railway tracks, there is no need to consider lane deviation or lane change [3, 5]. In addition, human-operated railway vehicles do not require obstacle detection by a camera

mounted on the train. But when the train is operated automatically at danger area, it is important to find quickly the obstacles near the tracks. Even if the brakes are in good condition, the train may hit the obstacles, if obstacles are not detected by cameras mounted on the train. As a danger area may be considered persons, animals and various vehicles that can enter the railway tracks. Therefore, a mobile device or a roadside unit separated from the train will probably be needed. These would be relay equipments to monitor the area near the tracks and railway crossings and send emergency information to the trains.

For railways, an automatic train operation controls the train inverter and braking devices by comparing their ground positions, Automatic Train Control (ATC) signals and speeds. The ATC is a control system that ensures safe train operation by comparing the allowable speed with the current train speed and automatically slowing down to the speed limit when the allowable speed is exceeded. In recent years, single-stage brake control has replaced multi-stage brake control based on the previous train location information. Single-stage brake control entails braking smoothly to the targeted stopping point. In the case of single-stage brake control, emergency brake such as obstacle avoidance may be difficult.

In many parts of the world, acceleration and deceleration operations on trains are performed primarily by the driver. When the strong g-force felt during deceleration, this makes it hard for people who are old or sick to keep their balance in the train car. This work focuses on deceleration assist control for railway vehicle drivers using Fuzzy Logic (FL). Intelligent solutions such as FL, machine learning, and deep learning make better decisions than humans in a number of different domains, including data science [4,7,12–14].

In our previous work [10,11], we proposed a Type-1 Fuzzy-based adaptive transmission control to reduce the computation cost and latency in wireless sensor networks. We used Type-1 Mamdani Fuzzy Inference and Type-1 Sugeno Fuzzy Inference. We considered not only observed time-series data but also 5 years of observed data imported from other research works.

In this paper, we propose a brake assisting control function based on FL. The proposed function provides an intelligent braking for the train drivers considering velocity, current brake level and environment status on the railway. We evaluate four Fuzzy inference types (models). The simulation results show that Type-2 FLC2 can provide better brake assisting function for train drivers than other models.

The structure of the paper is as follows. In Sect. 2, we describe the Fuzzy-based brake function. In Sect. 3, we explain the input and output parameters. In Sect. 4, we describe the evaluation results. Finally, conclusions and future work are given in Sect. 5.

2 Proposed Fuzzy-Based Brake Control Function

Fuzzy sets and FL [19] have been implemented to deal with vagueness and uncertainty in an inference process of an intelligent system such as knowledge-based system, logical control system and so on [2,6,8,9,15].

Both Mamdani Fuzzy inference model [9] and the Sugeno Fuzzy inference model [16] are classified as Type-1 inference model. We use a linear function for Sugeno Fuzzy inference model. In this way, we can reduce the computational complexity compared with the min-max-gravity model. We constructed a Type-1 Sugeno Fuzzy inference and Interval Type-2 Sugeno Fuzzy inference system written in MATLAB. The Interval Type-2 Membership Functions (MFs) constitute the upper and lower MFs. The Upper MF (UMF) is related to a Type-1 MF. For all conceivable input values, the Lower MF (LMF) is smaller than or equal to the UMF. The region between the UMF and the LMF is called the Footprint of Uncertainty (FOU). The proposed brake assist control function consists of two Fuzzy Logic Controllers (FLCs: FLC1 and FLC2). The controller basic functions are the fuzzifier, Sugeno Fuzzy inference engine, Fuzzy rule-base and defuzzifier. For input values, we use triangular and trapezoidal MFs for FLC1. While we use Gaussian and Sigmoidal MFs for FLC2. Then, for output values of both FLC1 and FLC2, we use a linear function, because they are good for real applications.

We consider the following FL models:

1. Type-1 Sugeno FLC1: Triangular and Trapezoidal MFs
2. Type-1 Sugeno FLC2: Gaussian and Sigmoidal MFs
3. Type-2 Sugeno FLC1: Triangular and Trapezoidal MFs
4. Type-2 Sugeno FLC2: Gaussian and Sigmoidal MFs

We use an Enhanced Iterative Algorithm with Stop Condition (EIASC) [18] as type-reduction method for converting a Type-2 output Fuzzy set to an Interval Type-1 Fuzzy set. We use three input parameters while the output parameter is the U. The Type-1 and Type-2 Fuzzy MFs of FLC1 and FLC2 are shown in Fig. 1 and Fig. 2, respectively. In Fig. 2, we show UMF (red line), LMF (blue line), and FOU (shaded area) for the Type-2 MFs.

3 Description of Input and Output Parameters

We explain in following Diff-Velocity, Brake and Env input parameters and U output parameter used for our Type-1 and Type-2 Sugeno Fuzzy inference models.

Diff-Velocity: The Diff-Velocity indicates the difference between the allowable velocity and the current velocity displayed in the cab of the train. We consider three levels of Diff-Velocity. If the allowable speed parameter is available at the train cab, a subsystem that reads the speed equipped with a camera module should be implemented.

Brake: We consider the current brake levels in the train. The Brake has three different levels, which can be interpreted as: Weak, Moderate (Mod) and Strong (Str).

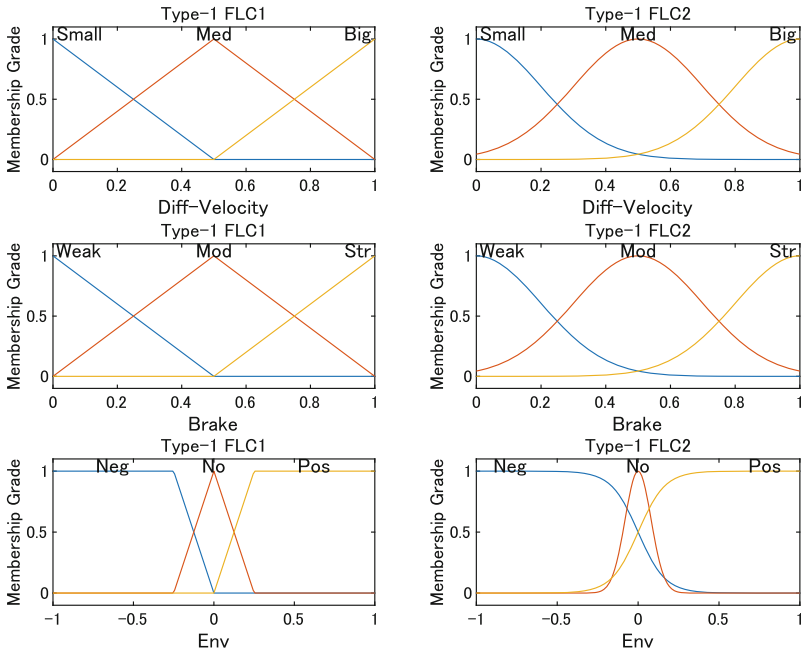


Fig. 1. Input MFs of Type-1 Sugeno Fuzzy for FLC1 and FLC2.

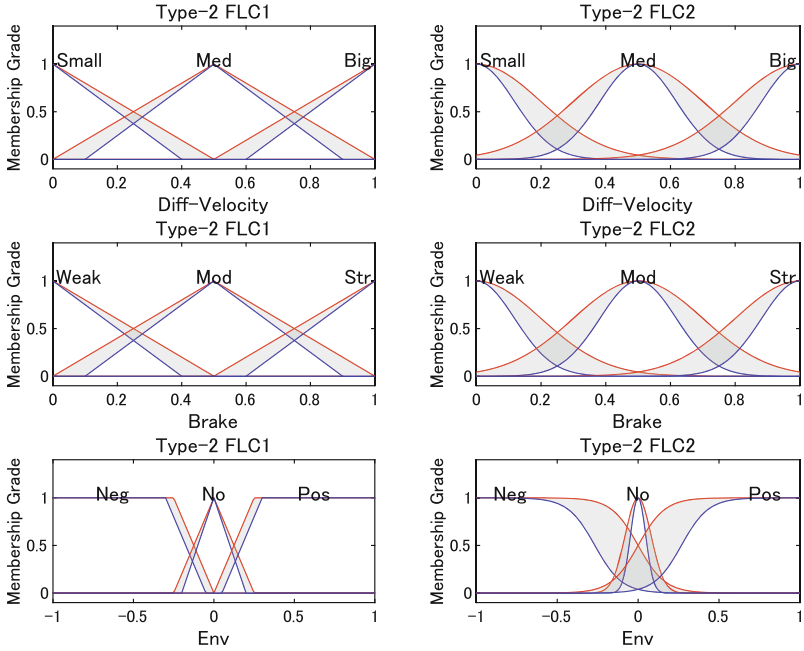


Fig. 2. Input MFs of Type-2 Sugeno Fuzzy for FLC1 and FLC2.

Table 1. Common rule-base of Type-1 and Type-2 FLCs.

Seq	Diff-Velocity	Brake	Env	U	Seq	Diff-Velocity	Brake	Env	U
1	Small	Weak	Neg	ExtrLow	15	Med	Mod	Pos	High
2	Small	Weak	No	ExtrLow	16	Big	Mod	Neg	Medium
3	Small	Weak	Pos	VeryLow	17	Big	Mod	No	High
4	Med	Weak	Neg	ExtrLow	18	Big	Mod	Pos	VeryHigh
5	Med	Weak	No	VeryLow	19	Small	Str	Neg	Low
6	Med	Weak	Pos	Low	20	Small	Str	No	Medium
7	Big	Weak	Neg	Low	21	Small	Str	Pos	High
8	Big	Weak	No	Medium	22	Med	Str	Neg	High
9	Big	Weak	Pos	High	23	Med	Str	No	VeryHigh
10	Small	Mod	Neg	VeryLow	24	Med	Str	Pos	ExtrHigh
11	Small	Mod	No	Low	25	Big	Str	Neg	VeryHigh
12	Small	Mod	Pos	Medium	26	Big	Str	No	ExtrHigh
13	Med	Mod	Neg	Low	27	Big	Str	Pos	ExtrHigh
14	Med	Mod	No	Medium					

Env: The Env indicates the environment status on the rail. To avoid accidents caused by natural disasters, it is important to make accurate assessments of the environment on a regular basis. In current railway system in Japan are installed rain sensors, anemometers and earthquake sensors. But these data are measured near railway tracks. In our approach, the environmental information is received in real time from sensors that are installed in the train. The value is between -1 and 1 . The Env has three different levels that can be interpreted as: Negative (Neg), Normal (No) and Positive (Pos).

U: Our system can control the U value in order to assist the train drivers. The U has seven different levels that can be interpreted as: Extremely-Low (ExtrLow), VeryLow, Low, Medium, High, VeryHigh and Extremely-High (ExtrHigh).

Table 1 shows the Fuzzy rule-base for matching the input and output parameters.

4 Evaluation Results

In this paper, we evaluate four FL models to assist the train drivers.

The simulation results are shown in Fig. 3. In this case, the Env parameter is 0, which represents a normal environment. Figure 3 shows the relationship between the input values Diff-Velocity and Brake, and the output U as a 3D surface graph. The colorbar indicates the level of the output value U.

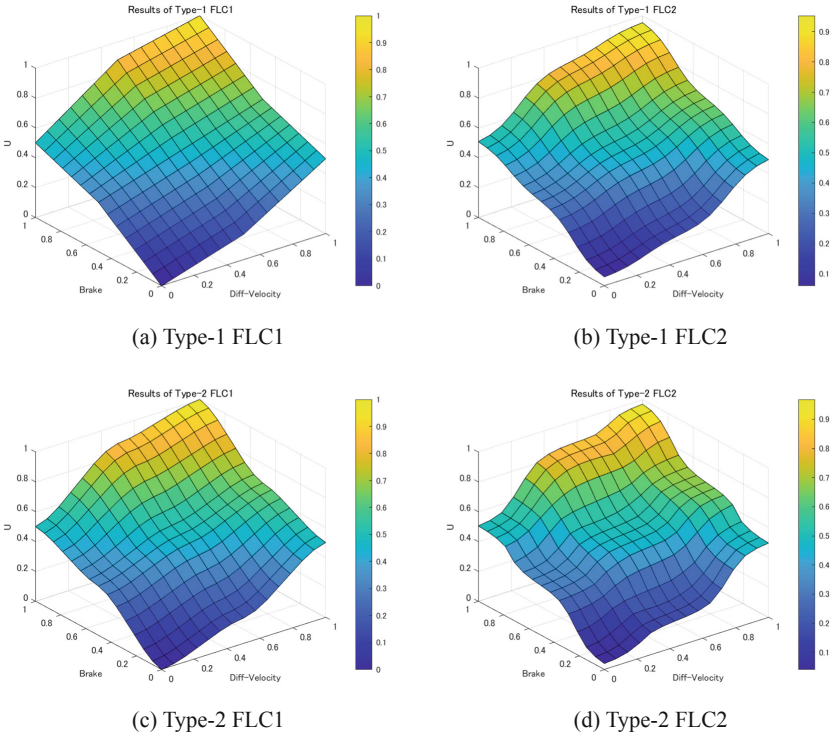


Fig. 3. Results for different models.

When we see the Brake axis, the Type-1 FLC1 is close to a straight line, while in Type-1 FLC2 the control is improved (see Fig. 3(a) and Fig. 3(b)). This is due to the effect of FLC2 MFs. Next, comparing Type-2 FLC1 and Type-2 FLC2, both Type-2 FLC2 models have better Brake control than Type-1 FLC1, but Type-2 FLC2 has smoother control than Type-2 FLC1 (see Fig. 3(c) and Fig. 3(d)).

We show the relationship between output U and Brake for different Diff-Velocity parameters from Fig. 4 to Fig. 6. The results of negative environment are shown in Fig. 4. For Type-1 FLC1 (Fig. 4(a)), until Diff-Velocity bounding is 0.6, we observe that the U is increased linearly with the increase of Brake value. If Diff-Velocity is more than 0.6, the slope increase is lower until Brake value is 0.5. For the other models from Fig. 4(b) to Fig. 4(d), the output value of U is controlled better.

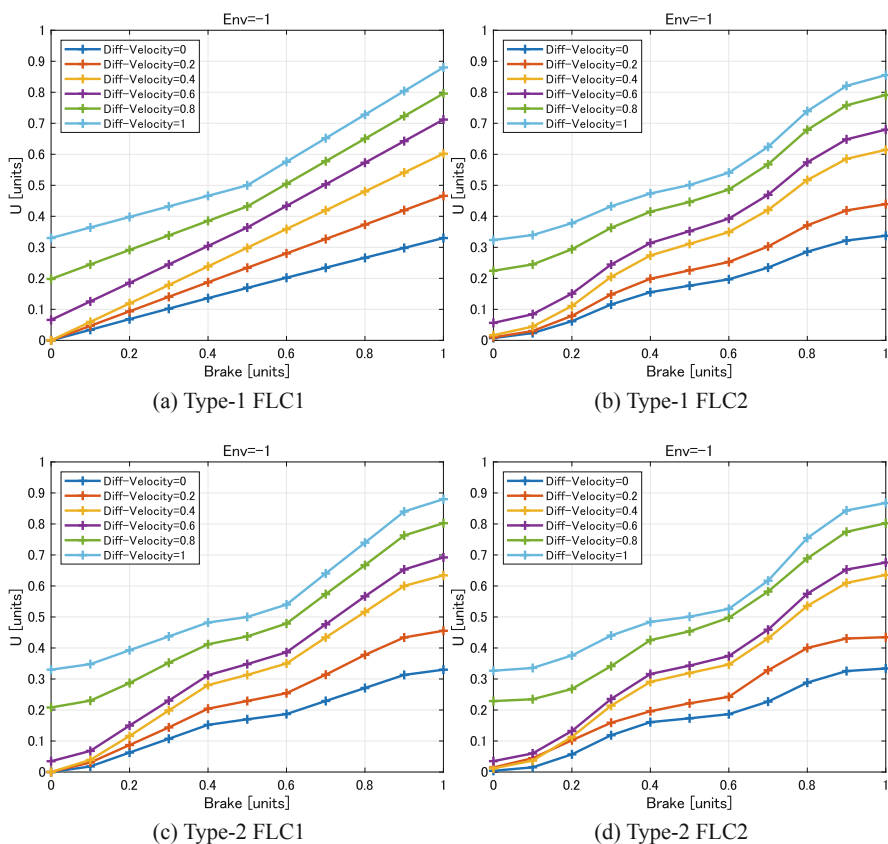


Fig. 4. Results for negative environment.

For normal and positive environments (see Fig. 5 and Fig. 6), the results of the U in Type-2 FLCs show a better stepwise braking control can be performed. Also, the output value of U is higher than the results of a negative environment. Thus, the proposed function provides mild brake force in negative situations and strong brake force in positive situations.

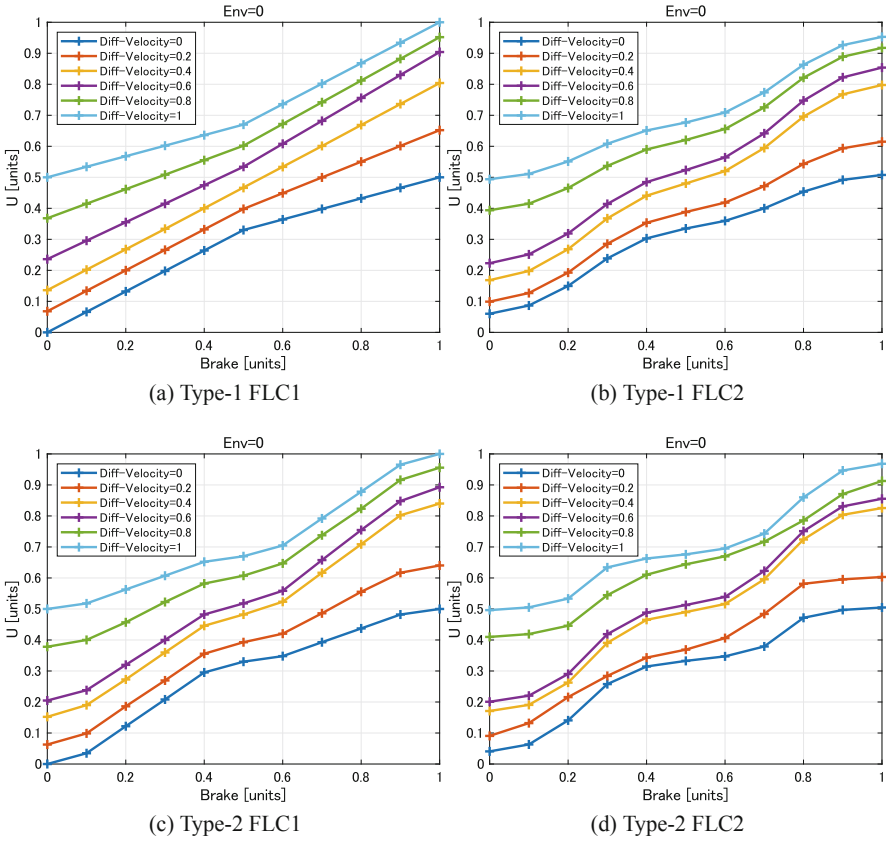


Fig. 5. Results for normal environment.

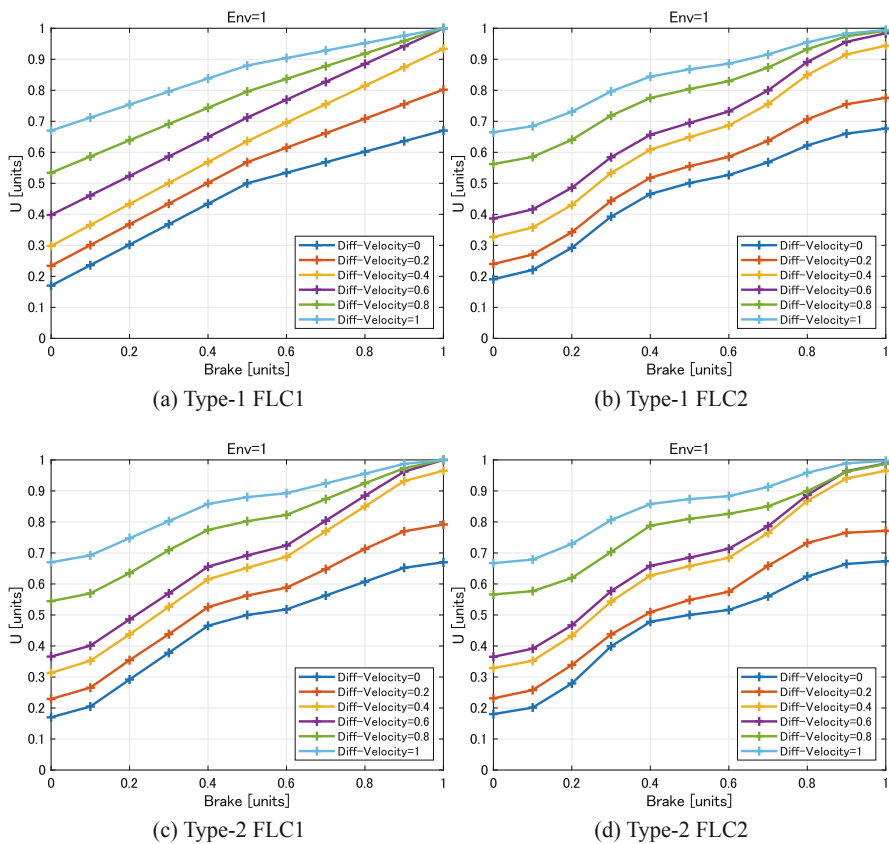


Fig. 6. Results for positive environment.

5 Conclusions

In this paper, we presented a brake assisting control function based on FL. The proposed function provides an intelligent breaking for the train drivers considering velocity, current brake level and environment status on the railway. We presented four FLC models. The simulation results show that Type-2 FLC2 can provide better brake assisting function for train drivers compared with other models.

In the future work, we will extend our simulation system considering other functions and other parameters

References

1. Asprilla, A.M., Martinez, W.H., Munoz, L.E., Cortes, C.A.: Design of an embedded hardware for motor control of a high performance electric vehicle. In: Proceedings of the IEEE Workshop on Power Electronics and Power Quality Applications (PEPQA-2017), pp. 1–5 (2017)
2. Balan, K., Manuel, M.P., Faied, M., Krishnan, M., Santora, M.: A fuzzy based accessibility model for disaster environment. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA-2019), pp. 2304–2310 (2019)
3. Cena, G., Cibrario Bertolotti, I., Hu, T., Valenzano, A.: On a software-defined CAN controller for embedded systems. *Comput. Stan. Interfaces* **63**, 43–51 (2019). <https://www.sciencedirect.com/science/article/pii/S0920548918302101>
4. Chimatapu, R., Hagra, H., Kern, M., Owusu, G.: Hybrid deep learning type-2 fuzzy logic systems for explainable AI. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE-2020), pp. 1–6 (2020)
5. Ghallabi, F., Nashashibi, F., El-Haj-Shhade, G., Mittet, M.A.: LIDAR-based lane marking detection for vehicle positioning in an HD map. In: Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC-2018), pp. 2209–2214 (2018)
6. Gupta, I., Riordan, D., Sampalli, S.: Cluster-head election using fuzzy logic for wireless sensor networks. In: Proceedings of the 3rd Annual Communication Networks and Services Research Conference (CNSR-2005), pp. 255–260 (2005)
7. Jammeh, E.A., Fleury, M., Wagner, C., Hagra, H., Ghanbari, M.: Interval type-2 fuzzy logic congestion control for video streaming across IP networks. *IEEE Trans. Fuzzy Syst.* **17**(5), 1123–1142 (2009)
8. Li, T.S., Chang, S.J., Tong, W.: Fuzzy target tracking control of autonomous mobile robots by using infrared sensors. *IEEE Trans. Fuzzy Syst.* **12**(4), 491–501 (2004)
9. Mamdani, E.H.: Application of fuzzy algorithms for control of simple dynamic plant. In: Proceedings of the Institution of Electrical Engineers, vol. 121, pp. 1585–1588 (1974)
10. Nishii, D., Ikeda, M., Barolli, L.: A fuzzy-based approach for reducing transmitted data considering data difference parameter in resilient WSNs. In: Barolli, L., Natwichai, J., Enokido, T. (eds.) EIDWT 2021. LNDECT, vol. 65, pp. 48–57. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70639-5_5
11. Nishii, D., Ikeda, M., Barolli, L.: A Takagi-Sugeno fuzzy-based adaptive transmission method in wireless sensor networks. In: Barolli, L. (ed.) 3PGCIC 2021. LNNS, vol. 343, pp. 279–288. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-89899-1_30
12. Petrakis, E.G.M., Sotiriadis, S., Soultanopoulos, T., Renta, P.T., Buyya, R., Bessis, N.: Internet of things as a service (iTaaS): Challenges and solutions for management of sensor data on the cloud and the fog. *Internet Things* **3–4**, 156–174 (2018)
13. Ruan, J., Jiang, H., Li, X., Shi, Y., Chan, F.T.S., Rao, W.: A granular GA-SVM predictor for big data in agricultural cyber-physical systems. *IEEE Trans. Industr. Inf.* **15**(12), 6510–6521 (2019)
14. Silver, D., et al.: Mastering the game of go without human knowledge. *Nature* **550**, 354–359 (2017)
15. Su, X., Wu, L., Shi, P.: Sensor networks with random link failures: distributed filtering for T-S fuzzy systems. *IEEE Trans. Industr. Inf.* **9**(3), 1739–1750 (2013)

16. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. In: IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, no. 1, pp. 116–132 (1985)
17. Wehner, P., Schwegelshohn, F., Gohringer, D., Hubner, M.: Development of driver assistance systems using virtual hardware-in-the-loop. In: Proceedings of the International Symposium on Integrated Circuits (ISIC-2014), pp. 380–383 (2014)
18. Wu, D., Nie, M.: Comparison and practical implementation of type-reduction algorithms for type-2 fuzzy sets and systems. In: Proceedings of the IEEE International Conference on Fuzzy Systems 2011 (FUZZ-IEEE 2011), pp. 2131–2138 (2011)
19. Zadeh, L.: Fuzzy logic, neural networks, and soft computing. *ACM Commun.* **37**, 77–84 (1994)



Preliminary Analysis of Performance Variation for ADS-B Position

Junichi Honda^(✉), Keisuke Matsunaga, Yasuyuki Kakubari, and Takuya Otsuyama

Electronic Navigation Research Institute,
National Research Institute of Maritime, Port, and Aviation Technology, Tokyo, Japan
j-honda@enri.go.jp

Abstract. This paper is concerned with the preliminary investigation of Automatic Dependent Surveillance – Broadcast (ADS-B) positional performance while the aircraft moves in the sky and on the ground. ADS-B broadcasts own position and other various operational information. Although each position is calculated by GNSS, positional accuracy changes depending on the flight situation and on-board devices. Blocking the GNSS signals and multipath immunity result in the performance degradation. Navigation integrity category (NIC) and navigation accuracy category-position (NACp) are provided as positional performance values of ADS-B. In this paper, analytical results of ADS-B regarding aircraft positions are shown. Statistical results of NIC and NACp on airborne and airport surface are demonstrated. In addition, focusing on an aircraft, the variation of performance values is presented. Causes of changing positional performance are discussed.

1 Introduction

Today, aircraft surveillance systems are important for supporting the safety aircraft operation. Surveillance systems based on a secondary surveillance radar (SSR) have become main aircraft surveillances in air traffic management. In these systems, a transponder loaded on an aircraft emits 1030/1090 MHz signals to reply to ground-based radar and for aircraft collision avoidance system. The frequency band of 1030 MHz is generally used for interrogation, and 1090 MHz is for reply or squitter. On the other hand, other surveillance systems were developed using the signal from the transponder. Multilateration (MLAT), wide-area MLAT (WAM), and Automatic Dependent Surveillance – Broadcast (ADS-B) were developed as application systems of SSR. They estimate the moving target positions by mainly using 1090 MHz signals.

ADS-B broadcasts own position calculated by GNSS, that is, a positional accuracy is based on the on-board GNSS receiver. The positional accuracy depends on the flight situation and on-board devices. Blocking the GNSS signals and multipath immunity result in the performance degradation. However, there is no way to detect the positional performance at ground receiver. Therefore, ADS-B emits some signal formats including the performance values. As performance parameters for aircraft position, a navigation integrity category (NIC) and a navigation accuracy category – position (NACp) can be extracted from ADS-B signals. Monitoring these values contributes the determination in the use of ADS-B whether the system is suitable for operational requirement.

We have started the evaluation of ADS-B performance in Japanese environment. In this study, statistics of ADS-B performance and variation of values will be discussed. This paper presents our system to collect ADS-B data, and then the procedure of evaluation is introduced. Preliminary experimental results will be shown. Focusing on an aircraft, in which situation the value changes is discussed.

2 ADS-B System

2.1 Mode S Format

There are two types of signals in 1090 MHz emitted by transponders [1] – one is an Air Traffic Control Beacon System (ATCRBS) and the other is a Mode Select (Mode S). ADS-B is transmitted on Mode S which provides the enhanced surveillance and communication capability required for air traffic control automation in comparison of ATCRBS. Mode S has not only surveillance position data also data link potential. The signal consists of twin pulse and data block parts in Fig. 1. Mode S message has 56 or 112-bits.

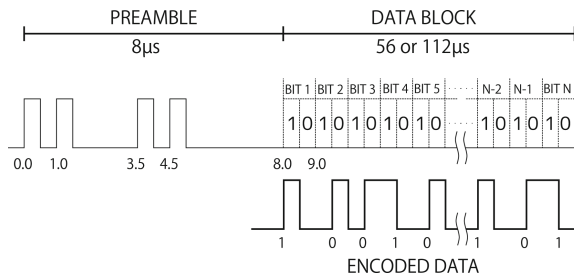


Fig. 1. Mode S format

2.2 ADS-B Format and Operational Status

Table 1 shows the representative reply and downlink formats (DF) of Mode S [2]. ADS-B is used in DF = 17. DF is determined by the first five bit of data block as illustrated in Fig. 2. CA is a transponder capability given by 3-bit, AA is 24-bit unique address assigned by International Civil Aviation Organization (ICAO), ME is 56-bit ADS-B message, and PI is 24-bit parity/interrogation field. Aircraft position and various operational information is included into ME (56-bit ADS-B message).

Information related to ADS-B position and operational status is made based on the data generated by GNSS receiver, and it broadcasts after inserting data on DF = 17 as shown in Fig. 3.

In this paper, ADS-B performance is discussed from a viewpoint of aircraft position. Currently three types of ADS-B versions exist. Versions 0 and 1 were earlier developed using extended squitter message, and Version 2 was developed to enhance integrity and

Table 1. Classification of DF

DF	Purpose
0	Short Air-Air Surveillance (ACAS)
4	Surveillance, Altitude
5	Surveillance, Identity
11	All-call
16	Long Air-Air Surveillance (TCAS)
17	Enhanced squitter (ADS-B)
18	Enhanced squitter transponder
20	Comm-B, Altitude
21	Comm-B, Identity

DF (10001)	CA: 3	AA: 24	ME: 56	PI: 24
---------------	-------	--------	--------	--------

Fig. 2. Format of DF = 17 for ADS-B

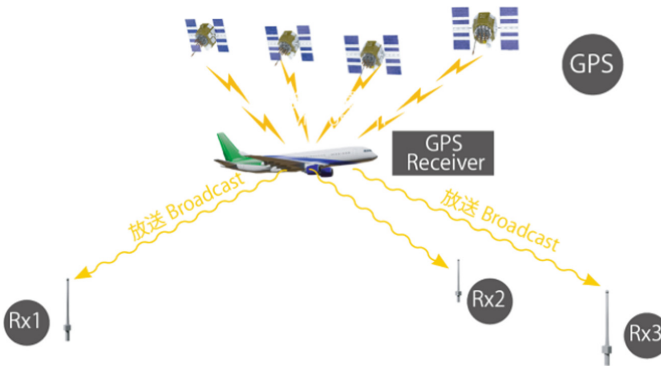


Fig. 3. Overview of ADS-B

accuracy reporting [3]. Parameters which are not covered in Versions 0 and 1, were included in Version 2. A new Version 3 was also issued in a latest document, but it doesn't work yet. To extract operational status, Version 2 or higher versions should be selected.

To evaluate ADS-B performance of aircraft position, ADS-B provides two values. One is NIC, and the other is NACp. NIC is an integrity value and allocated from 0 to 11. The value of NIC is determined by radius of containment (R_c) which is provided by the GNSS receiver. In addition, the provided number is different between airborne and surface movement. The relationship is summarized in Table 2 [3]. If an update has not received from an on-board data source within the past 2 s, R_c indicates unknown. NIC on

airport surface is also allocated to 0 and from 6 to 11. On the other hand, NACp provides the navigation accuracy, and the value is allocated from 0 to 11. The value from 12 to 15 is reserved currently. NACp is estimated by EPU (estimated position uncertainty) whose parameter is a 95% bound on horizontal position. If an update has not received from an on-board data source, NACp is also encoded as value indicating Unknown Accuracy. Allocation of NACp is shown in Table 3 [3]. NACp = 8 and 9 equivalent to position accuracy of general GPS. ICAO Circular 326 regarding the assessment of ADS-B defines both NIC and NACp values to meet operational performance [4]. In case of operation with 5 NM separation between aircraft, $NIC \geq 6$ and $NACp \geq 5$ should be satisfied.

Table 2. NIC encoding

NIC	Radius of containment (Rc)	Airborne			Surface		
		Airborne position TYPE code	NIC supplement codes		Surface position TYPE code	NIC supplement codes	
			A	B		A	C
0	Rc unknown	0, 18 or 22	0	0	0, 8	0	0
1	Rc < 20 NM (37.04 km)	17	0	0	N/A	N/A	N/A
2	Rc < 8 NM (14.816 km)	16	0	0	N/A	N/A	N/A
3	Rc < 4 NM (7.408 km)	16	1	1	N/A	N/A	N/A
4	Rc < 2 NM (3.704 km)	15	0	0	N/A	N/A	N/A
5	Rc < 1 NM (1852 m)	14	0	0	N/A	N/A	N/A
6	Rc < 0.6 NM (1111.2 m)	13	1	1	8	0	1
	Rc < 0.5 NM (926 m)	13	0	0	N/A	N/A	N/A
	Rc < 0.3 NM (555.6 m)	13	0	1	8	1	0
7	Rc < 0.2 NM (370.4 m)	12	0	0	8	1	1
8	Rc < 0.1 NM (185.2 m)	11	0	0	7	0	0
9	Rc < 75 m	11	1	1	7	1	0
10	Rc < 25 m	10 or 21	0	0	6	0	0
11	Rc < 7.5 m	9 or 20	0	0	5	0	0

Table 3. NACp encoding

NACp	Meaning = 95% horizontal accuracy bounds (EPU)
0	$EPU \geq 18.52 \text{ km}(10 \text{ NM})$ – Unknown accuracy
1	$EPU < 18.52 \text{ km}(10 \text{ NM})$ – RNP-10 accuracy
2	$EPU < 7.408 \text{ km}(4 \text{ NM})$ – RNP-4 accuracy
3	$EPU < 3.704 \text{ km}(2 \text{ NM})$ – RNP-2 accuracy
4	$EPU < 1852 \text{ m}(1 \text{ NM})$ – RNP-1 accuracy
5	$EPU < 926 \text{ m}(0.5 \text{ NM})$ – RNP-0.5 accuracy
6	$EPU < 555.6 \text{ m}(0.3 \text{ NM})$ – RNP-0.3 accuracy
7	$EPU < 185.2 \text{ m}(0.1 \text{ NM})$ – RNP-0.1 accuracy
8	$EPU < 92.6 \text{ m}(0.05 \text{ NM})$ – e.g. GPS (with SA)
9	$EPU < 30 \text{ m}$ – e.g. GPS (SA off)
10	$EPU < 10 \text{ m}$ – e.g. WAAS
11	$EPU < 3 \text{ m}$ – e.g. LAAS

Some signal formats are needed to determine values of ADS-B positional performance. Signal formats to analyze positional performance are shown in Table 4 [3]. There are four signal formats – airborne position (AP), airport surface position (SP), target state (TS) and status information (SI), and aircraft operational status (OS). Airborne position provides positional data in the sky, NIC and NIC supplement. Surface position provides positional data on the ground, and NIC. Target state and status information includes NACp. Aircraft operational status has ADS-B version, NACp and NIC supplement.

Table 4. Signal formats regarding ADS-B position

Classification	Maximum update interval
Extended Squitter Airborne Position (AP)	0.2 s
Extended Squitter Surface Position (SP)	0.2 s
Target State (TS) and Status Information (SI)	0.5 s
Extended Squitter Aircraft Operational Status (OS)	2.5 s

3 Measurement System

To observe ADS-B signals, a measurement system was deployed at Sendai Airport which is a midsize airport in Japan. ADS-B data have been collected for about three years. Figure 4 shows a schematic of data processing. Figure 4(a) is a picture of ADS-B receiver, and Fig. 4(b) is data flow. We selected ‘radarcape’ by jetvision corp. as ADS-B receiver. Mode S signals transmitted from aircraft transponders are detected and decoded by ADS-B receiver. Then the required DF = 17 signals are transmitted to a recorder. Signal processing unit carries out computing aircraft position and the value of performance.

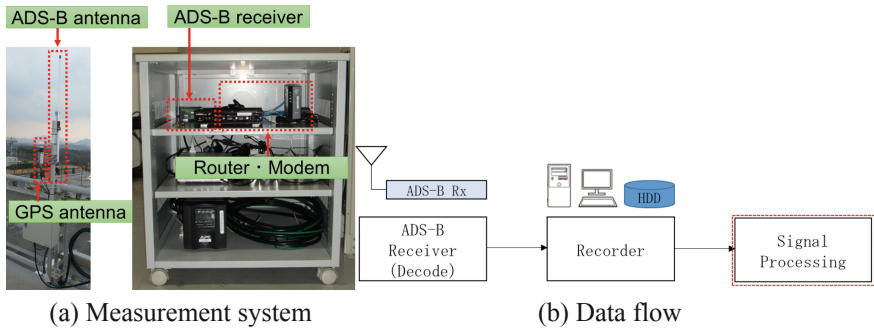


Fig. 4. Schematic of data processing

Figure 5 illustrates the procedure for estimating ADS-B positional performance. Abbreviation shown in Fig. 5 was defined in Table 4. Firstly, ADS-B version is determined by the aircraft operational status for each aircraft with 24-bit ICAO address. Only Version 2 aircraft is picked up for NIC and NAC_p analysis. Then, the value of NIC supplement is extracted from airborne position and the target state and status information. For aircraft in the sky, NIC supplements A and B are extracted from airborne position message and aircraft operational status message, respectively. For aircraft on airport surface, NIC supplements A and C are extracted from the aircraft status information message. These values are stored and updated to obtain NIC values. Next, we classified the data into two categories based on the airborne position and the surface position. Format type code in airborne positions is combined with NIC supplements A and B, and format type code in surface positions is combined with NIC supplements A and C. Finally, NIC value is defined for an airborne position or a surface position. On the other hand, NAC_p for an airborne position or a surface position is easily determined by target state and status information, and aircraft operational status. We analyze statistic of NIC and NAC_p values, and estimate performance variation in comparison of target position.

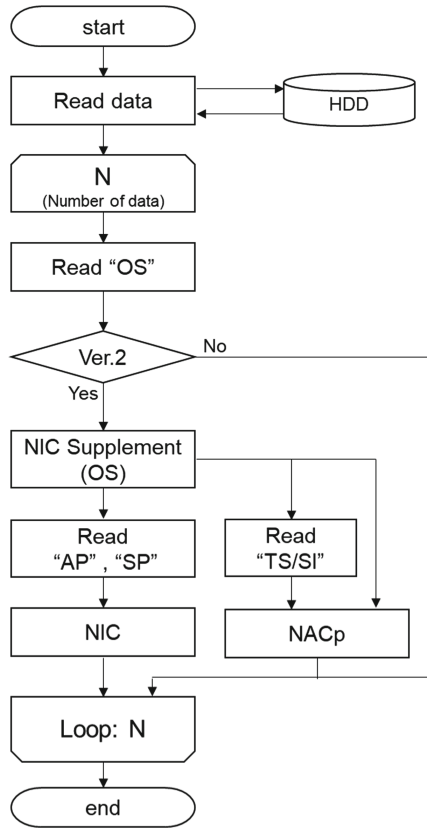


Fig. 5. Determination procedure of ASD-B performance

4 Measurement Result

Statistical analysis results, and performance variation vs. elapsed time are shown in this section. We collected ADS-B data for a month in Sendai Airport in May 2022. Data of more than 2000 mode S aircraft were processed.

Firstly, we analyzed the number of aircraft for each ADS-B version. The number was 582 for Version 0, 8 for Version 1, and 1632 for Version 2. It is found that the rate of latest Version 2 was about 73%, and second largest number of versions was Version 0 whose rate was about 26% during the data collection period.

Next, NIC values were evaluated. Table 5 shows the number of position data with NIC value. NIC = 8 dominates over 99% which is satisfied with the operation for 3 NM and 5 NM separations. We can see NIC = 0 data which means Rc.

Table 5. Number of NIC values

NIC	0	1	2	3	4	5
Num.	4547	27	38	0	0	40
Rate (%)	0.1	7e-5	1e-4	0	0	1e-4
NIC	6	7	8	9	10	11
Num.	740	80380	35078531	60244	47394	0
Rate (%)	0.002	0.22	99.4	0.17	0.13	0

Table 6 shows the number of position data with each NACp value for Version 2. As described in Sect. 2, NACp represents horizontal positional error based on a 95% bound. In the analysis of data for a month, the largest number of NACp was 9 whose error is within 30 m. Rate of position data with NACp = 9 and more indicates over 99%. It is found from results of NIC and NACp status that signals over 90% has the capability of high integrity and high positional accuracy.

Table 6. Number of NACp values

NACp	0	1	2	3	4	5
Num.	2322	0	0	0	0	0
Rate (%)	0.006	0	0	0	0	0
NACp	6	7	8	9	10	11
Num.	21	76	220692	26148808	8572929	5845
Rate (%)	6e-5	0.0002	0.63	74.8	24.5	0.016

Finally, performance variation is discussed. There are mainly three reasons that performance parameters change: one is the shadowing of the signals from satellites, second is the multipath interferences caused by buildings, and third is the something trouble of GNSS receivers such as instant shutdown. Figure 6 shows the NIC values vs. measurement time. We plotted NIC values for each aircraft during the flight. ADS-B data of 1632 aircraft for a month was divided for each day and were processed. Each data is illustrated as the first signal observed time becomes zero second. In this figure, several errors for elapsed time exist, but the reason of the error is not discussed in this paper. The figure demonstrates that NIC values change from 8 to 6 or 7. Figure 7 is the enlarged figure of Fig. 6 when elapsed time is in range from 0 to 200 s. The number of aircraft with NIC variation was 97. Minimum parameter changes were 2 times, and maximum ones were 7 times. Table 7 shows the number of aircraft with NIC value changes. In this paper, an aircraft is picked up to evaluate the phenomenon of positional parameter variation.

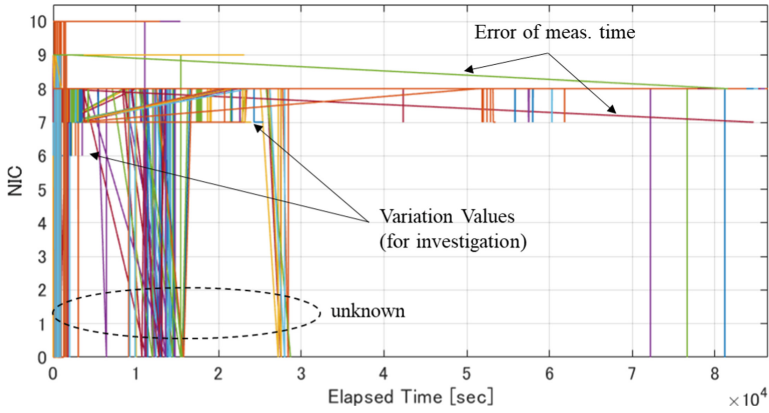


Fig. 6. Values of NIC vs. elapsed time

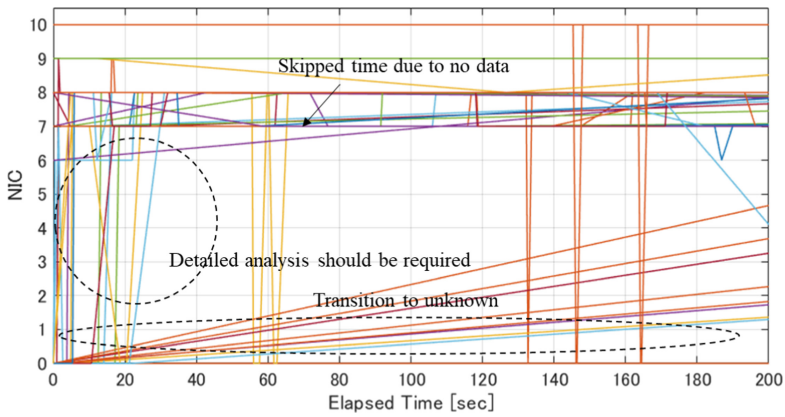


Fig. 7. Values of NIC vs. elapsed time in range of 0 to 200 s

Table 7. Number of aircraft which has several NIC values

Variation times	2	3	4	5	6	7
Number	81	10	0	3	2	1

Focusing on some of the NIC variations, we confirmed aircraft routes. Figure 8 shows thirteen trajectories of an aircraft (Fig. 8(a)) and one trajectory with NIC variation (Fig. 8(b)). In Fig. 8(b), NIC values are expressed by different color - yellow is NIC = 8 and blue is NIC = 0. An aircraft that NIC value changes seven times shown in Table 7, was selected. In Fig. 8(a), each trajectory painted by different colors. These aircraft flew over the ocean, and those thirteen trajectories were almost coincided. However, it is found that one trajectory obtained variation of NIC values. Worse NIC values were observed at the right part in Fig. 8(b). In this case, since an aircraft was over oceans,

multipath interferences was not the cause. Other twelve trajectories have almost $NIC = 8$. Therefore, it is considered that the GNSS receiver/satellites were out of service instantly or an ionosphere effect was observed.

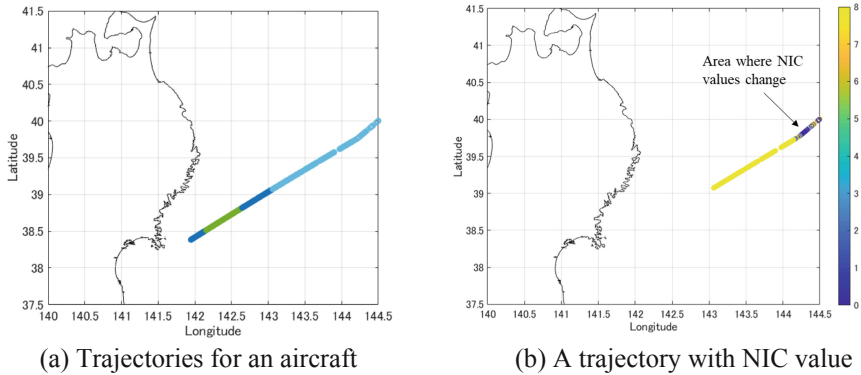


Fig. 8. Trajectories for some days and a trajectory with NIC variation

5 Conclusion

In this paper, we discussed the performance variation of ADS-B. Firstly, the principle of ADS-B and performance parameters regarding aircraft position were introduced. NIC and NACp of position data were obtained by using information in ADS-B message. Then, measurement system was presented. ADS-B data observed for a month was analyzed. Number of the NIC and NACp values were shown. $NIC = 8$ and $NACp \geq 9$ dominated in the total signals. It was found that the almost signals provide high integrity and high accuracy which satisfy the requirements for operation with 5 NM separation. However, in the analysis of NIC vs. elapsed time, we obtained the value's variation at some aircraft. Focusing on an aircraft with NIC variation, a trajectory with NIC values was illustrated. It was found that the cause other than multipath interference was observed. Additional analysis is needed.

In the near future, we would like to analyze the ADS-B data on airport surface in detail and clarify how the performance degradation is generated.

References

1. RTCA, Inc.: Minimum Operational Performance Standards for Air Traffic Control Radar Beacon System/Mode Select (ATCRBS/Mode S) Airborne Equipment, RTCA DO-181E (2011)
2. Aeronautical Telecommunications Annex 10 Volume IV Surveillance and Collision Avoidance System, 4th edition, ICAO (International Civil Aviation Organization), QC, Canada, July (2007)

3. Technical Provisions for Mode S Services and Extended Squitter, Doc 9871 Second Edition, ICAO (International Civil Aviation Organization), QC, Canada (2012)
4. Assessment of ADS-B and Multilateration Surveillance to Support Air Traffic Services and Guidelines for Implementation, Cir 326, ICAO (International Civil Aviation Organization), QC, Canada (2012)



A Simulation System for Mobility Control of Swarm Drones to Provide Wireless Mesh Network Services

Yuma Yamashita¹, Nobuki Saito², Chihiro Yukawa², Kyohei Toyoshima², Tetsuya Oda¹✉, Kengo Katayama¹, and Leonard Barolli³

¹ Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
t20j091yy@ous.jp, oda@ous.ac.jp, katayama@ice.ous.ac.jp

² Graduate School of Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
{t21jm01md,t22jm19st,t22jm24jd}@ous.jp

³ Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan
barolli@fit.ac.jp

Abstract. Wireless Mesh Networks (WMNs) provide a stable network over a wide area by configuring the network like a mesh. In order to provide a lower cost and more stable network, various methods for optimizing the placement of mesh routers are being studied. In this paper, we propose a simulation system for mobility control of swarm drones based on DQN to provide WMNs services. Also, we present the simulation results considering the normal distribution of mesh clients. The simulation results show that the proposed system can control the mobility of swarm drones and provide WMNs services.

1 Introduction

The Wireless Mesh Networks (WMNs) [1–3] provide a stable network over a wide area by configuring the network like a mesh. In order to provide a lower cost and more stable network, various methods for optimizing the placement of mesh routers are being studied. In our previous work [4–6], we proposed and evaluated different meta-heuristics such as Genetic Algorithms (GA) [7], Hill Climbing (HC) [8], Simulated Annealing (SA) [9], Tabu Search (TS) [10] and Particle Swarm Optimization (PSO) [11] for mesh router placement optimization.

Many research works have been done in recent years to provide communication infrastructure for drone, hot air balloons or satellites. In order to cover a large communication area, a small number of units are needed, but the cost of operation is high. Also, the units such as hot air balloons cannot even be launched if the climate situation in the sky is not stable. Autonomous Aerial Vehicles (AAVs) [12] are types of drones that have the ability to operate autonomously without human control and are expected to be used in a variety of fields.

In this work, we consider small drones with low launching and operating costs to provide communication areas for users that want to use WMNs service. In [13–17] the authors consider Wireless Sensor and Actuator Networks (WSANs), which can act autonomously for disaster monitoring. A WSAN consists of wireless network nodes, all of which have the ability to sense events (sensors) and perform actuation (actuators) based on the sensing data collected by the sensors. WSAN nodes in these applications are nodes with integrated sensors and actuators that have high processing power, high communication capability, high battery capacity and may include other functions such as mobility. The application areas of WSAN include AAV [18], Autonomous Surface Vehicle (ASV) [19], Heating, Ventilation, Air Conditioning (HVAC) [20], Internet of Things (IoT) [21], Ambient Intelligence (AmI) [22], ubiquitous robotics [23], and so on.

In this paper, we propose a simulation system for mobility control of swarm drones based on DQN to provide WMNs services. Also, we present the simulation results considering the normal distribution of mesh clients.

The structure of the paper is as follows. In Sect. 2, we give a short description of mesh router placement problem. In Sect. 3 presents a brief introduction of DQN. In Sect. 4, we present the proposed simulation system. In Sect. 5, we discuss the simulation results. Finally, conclusions and future work are given in Sect. 6.

2 Mesh Router Nodes Placement Problem

We consider multiple mesh router nodes and arbitrarily distributed mesh client nodes in a two-dimensional contiguous region [24]. The objective is to place mesh router nodes in the considered area in such way that the network connectivity and the number of mesh client node coverage are maximized.

The connected graph of mesh router nodes is considered a component, with the largest component being the Giant Component (GC). Since the communication range is maximum when all components are linked, the Size of GC (SGC) is used as a measure of network connectivity. The Number of Covered Mesh Client (NCCM) is the number of covered mesh client nodes within the communication range of the GC.

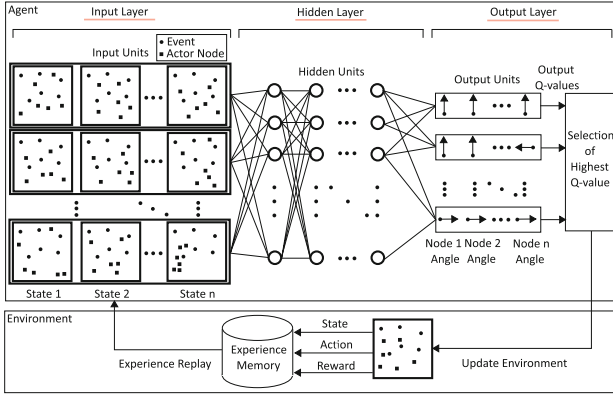


Fig. 1. The structure of DQN based simulation system for mobility control of swarm drones.

3 DQN

Deep reinforcement learning is a function approximation method using deep neural network, which can be used for controlling autonomous robots such as AAVs. Deep Q-Network (DQN), which uses a Convolution Neural Network (CNN) as a function approximation of Q-learning is a Deep Reinforcement Learning method proposed by Mnih et al. [25,26]. In paper [25], the authors present the implementation and performance evaluation of DQN for different Atari 2600 games. The authors show that DQN can use game screen input without feature value design, and they have better performance than conventional reinforcement learning method with feature value design using linear function [27]. The DQN combines the methods of neural fitted Q iteration [28], experience replay [29], sharing the hidden layer of action value function in each behavior pattern, and learning can be stabilized even with a nonlinear function such as CNN [30,31].

Figure 1 shows the structure of the DQN based simulation system. In this work, we use the Deep Belief Network (DBN), where computational complexity is smaller than CNN for DNN part in DQN. The environment is set as v_i . At each step, the system selects an action from the action sets of the mobile actor nodes and observes the communication coverage v_t from the current state. The change of the mobile actor node score r_t is regarded as the reward for the action. For reinforcement learning, we can consider all of these mobile actor nodes sequences m_t as Markov decision process [32], where sequences of observations and actions are $m_t = v_1, a_1, v_2, \dots, a_{t-1}, v_t$. It uses a method known as experience replay in which it stores experiences at each timestep, $e_t = (m_t, a_t, r_t, m_{t+1})$ in a dataset $D = e_1, \dots, e_N$, cached over many episodes into a Experience Memory. Defining the discounted reward for the future by a factor γ , the sum of the future reward until the end would be $R_t = \sum_{t'=t}^T \gamma^{t'-t} r_{t'}$. T means the termination time-step of the mobile actor nodes.

Algorithm 1. Derive reward for an agent.

Require: Derive *SGC*, *NCCM* from the location of the actor node and mesh clients.

Ensure: cooperative reward

```

1: reward ← 0
2: if Covered any mesh clients then
3:   reward+ = Number of Covered Mesh Clients by an Agent
4: else
5:   reward− = Average(Number of Covered Mesh Clients by an Agent)
6: if Belongs to the giant component then
7:   reward+ = Number of Covered Mesh Clients by Belong Component
8: else
9:   reward− = Number of Covered Mesh Clients by Belong Component
10: if Overlapped with other actor nodes then
11:   reward− =  $\frac{\text{Number of Covered Mesh Clients by an Agent}}{2}$ 

```

After running experience replay, the system selects and executes an action according to an ϵ -greedy strategy. Since using histories of arbitrary length as inputs to a neural network can be difficult, Q function instead works on fixed length format of histories produced by a function ϕ . The target was to maximize the action value function $Q^*(m, a) = \max_{\pi} E[R_t | m_t = m, a_t = a, \pi]$, where π is the strategy for selecting of best action. From the Bellman equation, it is equal to maximize the expected value of $r + \gamma Q^*(m', a')$, if the optimal value $Q^*(m', a')$ of the sequence at the next time step is known.

$$Q^*(m', a') = E_{m' \sim \xi} [r + \gamma \max_a Q^*(m', a') | m, a] \quad (1)$$

By not using iterative updating method to optimize the equation, it is common to estimate the equation by using a function approximator. Q-network in DQN is a neural network function approximator with weights θ and $Q(s, a; \theta) \approx Q^*(m, a)$. The loss function to train the Q-network is:

$$L_i(\theta_i) = E_{s, a \sim \rho(\cdot)} [(y_i - Q(s, a; \theta_i))^2]. \quad (2)$$

The y_i is the target, which is calculated by the previous iteration result θ_{i-1} . $\rho(m, a)$ is the probability distribution of sequences m and a . The gradient of the loss function is shown in Eq. (3):

$$\nabla_{\theta_i} L_i(\theta_i) = E_{m, a \sim \rho(\cdot); s' \sim \xi} [(y_i - Q(m, a; \theta_i)) \nabla_{\theta_i} Q(m, a; \theta_i)]. \quad (3)$$

4 Proposed Method

In this section, we present the design and implementation of the proposed simulation system based on DQN for mobility control of swarm drones as actor nodes in WSAW to provide WMNs services. The actor node can choose the networking, moving direction, actuation and sensing policies to maximize NCCM and SGC.

Table 1. Simulation parameters of DQN.

Parameters	Values
Number of episode	10000
Number of iteration	60
Number of hidden layers	2
Number of hidden units	64
Initial weight value	Normal Initialization
Activation function	Sigmoid
Action selection probability (ϵ)	$0.999 - (t / \text{Number of episode})$ ($t = 0, 1, 2, \dots, \text{Number of episode}$)
Learning rate (α)	0.04
Discount rate (γ)	0.5
Experience memory size	500×100
Batch size	32
Number of actor nodes	4
Number of mesh clients	24
Initial placement of actor nodes $[X, Y]$	$[16.0, 16.0]$
Initial placement of mesh clients	Normal distribution
Area size $[(X_{min}, X_{max}), (Y_{min}, Y_{max})]$	$[(0.0, 32.0), (0.0, 32.0)]$
Coverage of actor node	3.2

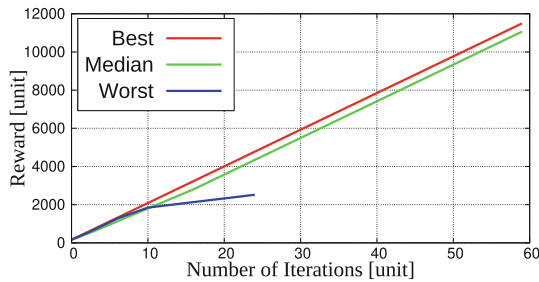
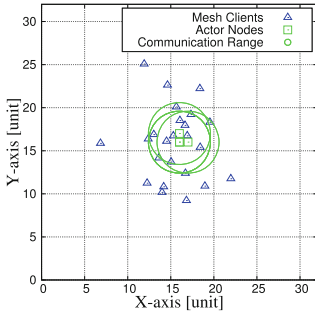


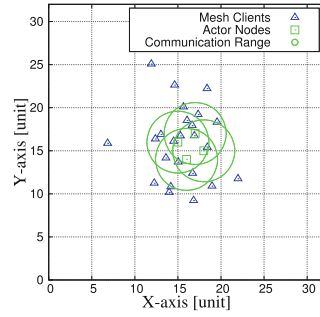
Fig. 2. Simulation results of rewards.

We consider tasks in which nodes interact with other nodes in an environment. In this case, the actor node moves step by step in a sequence of observations, actions and rewards. We took into consideration the connectivity and mobility of actor nodes. For an actor node are considered 5 mobile patterns (back, forward, right, left, stop). The actor nodes have networking, sensing, mobility and actuation mechanisms. In order to decide the reward function, we considered GC) and NCMC parameters. The reward function r is defined in Algorithm 1.

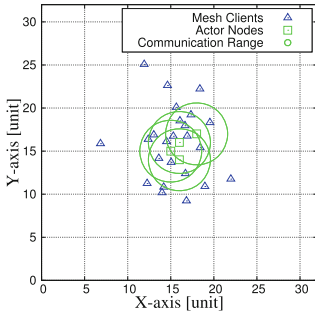
The proposed system based on DQN enables WMNs services to provide a large communication area with drones by maximizing the rewards. The initial



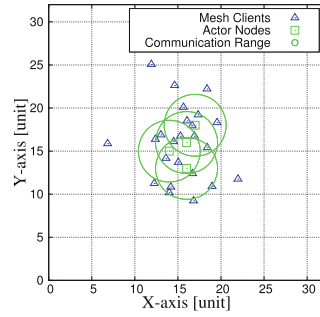
(a) Number of iteration = 1.



(b) Number of iteration = 20.



(c) Number of iteration = 40.



(d) Number of iteration = 60.

Fig. 3. Visualization results of best episode.

weights values are assigned as Normal Initialization [33]. The input layer is using actor nodes and the position of events, total reward values in Experience Memory and mobile actor node patterns. The hidden layer uses a Sigmoid function to learn negative rewards. The output Q values are actor node movement patterns.

5 Simulation Results

In this section, we discuss the simulation results for mobility control of swarm drones to provide WMNs services. Table 1 shows the parameters used in the simulations. For simulations, we consider 24 mesh clients, normal distributions and 4 actor nodes.

Figure 2 shows the change in total reward value of the action in each iteration for **Worst**, **Median**, and **Best** episodes in DQN. We can see that with higher reward values, the actor node can move and also keep a connection with other actor nodes. In the case of **Worst** episode reward, the actor node moved out of the target area and the episode is interrupted.

Figure 3 shows the visualization results for the movement of the **Best** episodes in DQN when the number of iteration is 1, 20, 40 and 60. From the visualization

results, we see that the actor node moves to cover many mesh clients. The proposed system can control the mobility of swarm drones and provide WMNs services to mesh clients.

6 Conclusions

In this paper, we propose a simulation system for mobility control of swarm drones based on DQN to provide WMNs services. From the visualization results, we found that the actor node moves in such a way in order to cover many clients. In addition, the proposed system can control the mobility of swarm drones and provide WMNs services.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number JP20K19793.

References

1. Akyildiz, I.F., et al.: Wireless mesh networks: a survey. *Comput. Netw.* **47**(4), 445–487 (2005)
2. Jun, J., et al.: The nominal capacity of wireless mesh networks. *IEEE Wirel. Commun.* **10**(5), 8–15 (2003)
3. Oyman, O., et al.: Multihop relaying for broadband wireless mesh networks: from theory to practice. *IEEE Commun. Mag.* **45**(11), 116–122 (2007)
4. Oda, T., et al.: WMN-GA: a simulation system for wmnns and its evaluation considering selection operators. *J. Ambient. Intell. Humaniz. Comput.* **4**(3), 323–330 (2013). <https://doi.org/10.1007/s12652-011-0099-2>
5. Ikeda, M., et. al.: Analysis of WMN-GA simulation results: WMNs performance considering stationary and mobile scenarios. In: *Proceedings of The 28-th IEEE International Conference on Advanced Information Networking and Applications (IEEE AINA-2014)*, pp. 337-342 (2014)
6. Oda, T., Elmazi, D., Barolli, A., Sakamoto, S., Barolli, L., Xhafa, F.: A genetic algorithm-based system for wireless mesh networks: analysis of system data considering different routing protocols and architectures. *Soft Comput.* **20**(7), 2627–2640 (2015). <https://doi.org/10.1007/s00500-015-1663-z>
7. Holland, J.H.: Genetic algorithms. *Sci. Am.* **267**(1), 66–73 (1992)
8. Skalak, D.B.: Prototype and feature selection by sampling and random mutation hill climbing algorithms. In: *Proceedings of The 11-th International Conference on Machine Learning (ICML-1994)*, pp. 293–301 (1994)
9. Kirkpatrick, S., et al.: Optimization by simulated annealing. *Science* **220**(4598), 671–680 (1983)
10. Glover, F.: Tabu search: a tutorial. *Interfaces* **20**(4), 74–94 (1990)
11. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: *Proceedings of The IEEE International Conference on Neural Networks (ICNN-1995)*, pp. 1942-1948 (1995)
12. Stöcker, C., et al.: Review of the current state of UAV regulations. *Remote Sens.* **9**(5), 1–26 (2017)

13. Oda, T., et al.: Design and implementation of a simulation system based on deep Q-network for mobile actor node control in wireless sensor and actor networks. In: Proceedings of The 31-th IEEE International Conference on Advanced Information Networking and Applications Workshops (IEEE AINA-2017), pp. 195-200 (2017)
14. Oda, T., Elmazi, D., Cuka, M., Kulla, E., Ikeda, M., Barolli, L.: Performance evaluation of a deep Q-network based simulation system for actor node mobility control in wireless sensor and actor networks considering three-dimensional environment. In: Barolli, L., Woungang, I., Hussain, O.K. (eds.) INCoS 2017. LNDECT, vol. 8, pp. 41–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-65636-6_4
15. Oda, T., Kulla, E., Katayama, K., Ikeda, M., Barolli, L.: A deep Q-network based simulation system for actor node mobility control in WSANs considering three-dimensional environment: a comparison study for normal and uniform distributions. In: Barolli, L., Javaid, N., Ikeda, M., Takizawa, M. (eds.) CISIS 2018. AISC, vol. 772, pp. 842–852. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-93659-8_77
16. Saito, N., et al.: A Tabu list strategy based DQN for AAV mobility in indoor single-path environment: implementation and performance evaluation. *Internet Things* **14**, 100394 (2021)
17. Saito, N., Oda, T., Hirata, A., Yukawa, C., Kulla, E., Barolli, L.: A LiDAR based mobile area decision method for TLS-DQN: improving control for AAV mobility. In: Barolli, L. (ed.) 3PGCIC 2021. LNNS, vol. 343, pp. 30–42. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-89899-1_4
18. Sandino, J., et al.: UAV framework for autonomous onboard navigation and people/object detection in cluttered indoor environments. *Remote Sens.* **12**(20), 1–31 (2020)
19. Moulton, J., et al.: An autonomous surface vehicle for long term operations. In: Proceedings of MTS/IEEE OCEANS, pp. 1-10 (2018)
20. Oda, T., Ueda, C., Ozaki, R., Katayama, K.: Design of a deep Q-Network based simulation system for actuation decision in ambient intelligence. In: Barolli, L., Takizawa, M., Xhafa, F., Enokido, T. (eds.) WAINA 2019. AISC, vol. 927, pp. 362–370. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15035-8_34
21. Oda, T., et al.: Design and implementation of an IoT-based e-learning testbed. *Int. J. Web Grid Serv.* **13**(2), 228–241 (2017)
22. Hirota, Y., Oda, T., Saito, N., Hirata, A., Hirota, M., Katatama, K.: Proposal and experimental results of an ambient intelligence for training on soldering iron holding. In: Barolli, L., Takizawa, M., Enokido, T., Chen, H.-C., Matsuo, K. (eds.) BWCCA 2020. LNNS, vol. 159, pp. 444–453. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61108-8_44
23. Hayosh, D.: Woody: low-cost, open-source humanoid torso robot. In: Proceedings of The 17-th International Conference on Ubiquitous Robots (ICUR-2020), pp. 247-252 (2020)
24. Oda, T., et al.: Evaluation of WMN-GA for different mutation operators. *Int. J. Space Based Situated Comput.* **2**(3), 149–157 (2012)
25. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)
26. Mnih, V., et. al.: Playing atari with deep reinforcement learning. In: [arXiv:1312.5602v1](https://arxiv.org/abs/1312.5602v1), pp. 1-9 (2013)
27. Lei, T., Ming, L.: A robot exploration strategy based on q-learning network. In: Proceedings of IEEE International Conference on Real-time Computing and Robotics (RCAR-2016), pp. 57-62 (2016)

28. Riedmiller, M.: Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 317–328. Springer, Heidelberg (2005). https://doi.org/10.1007/11564096_32
29. Lin, L.J.: reinforcement learning for robots using neural networks. Technical report, DTIC Document (1993)
30. Lange, S., Riedmiller, M.: Deep auto-encoder neural networks in reinforcement learning. In: Proceedings of The 2010 International Joint Conference on Neural Networks (IJCNN-2010), pp. 1-8(2010)
31. Kaelbling, L.P., et al.: Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**(1–2), 99–134 (1998)
32. Kaelbling, L.P., et al.: Reinforcement learning: a survey. *J. Artif. Intell. Res.* **4**, 237–285 (1996)
33. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13-th International Conference on Artificial Intelligence and Statistics (AISTATS-2010), pp. 249-256 (2010)



Comparison of Transmission Spectra of Fork-Shaped Photonic Crystal Branch Waveguide for Continuous and Band-Limited Input Signal

Hiroshi Maeda^(✉)

Department of Information and Communication Engineering,
Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Fukuoka 811-0295, Japan
hiroshi@fit.ac.jp

Abstract. Photonic crystal waveguide with fork-shaped three branch outputs were numerically analyzed by finite difference time domain (FDTD) method. The transmission spectra of fork-shaped photonic crystal waveguide for continuous and band-limited impulse input signal are compared. The transmission spectra showed typical four wavelengths with obvious high output power peaks at center waveguide of the fork-shaped branch.

1 Introduction

Photonic crystal structures (PhCs) or electromagnetic band gap (EBG) structures have periodic distribution of material constants and are applied into practical use in optical components for signal generation, transmission and reception, because of its unique and sensitive characteristics with respect to the signal frequency. Those characteristics are based on photonic band gap (PBG) phenomena [1–4]. In signal procession and transmission utilizing PBG devices in optical integrated circuits, high density multiplexing in frequency domain is expected due to its sensitivity with respect to optical wavelength. This is important to improve capacity of information transmission in photonic network with dense multiplexing technique of signal in wavelength domain.

The behavior of electromagnetic wave in periodic structure can be controlled by selecting material constants, designing periodic profile of the structure and the frequency spectrum range of the signal. For various kinds of materials and for various frequency ranges of application purposes, PBG might be found by designing the structure with fundamental unit lattice. This means that, by setting the parameters appropriately, confinement and transmission of electromagnetic wave along line-defect in the structure is possible for desired range of frequency from microwave to optical domain. In this meaning, we examined the propagation characteristics of two dimensional photonic crystal waveguide with fork-shaped branch with square lattice of dielectric pillar in optical frequency range. In author's previous works [5–14], authors have numerically demonstrated

wave propagation in the photonic crystal waveguide in optical frequency domain by using silica rods as dielectric pillar. Additionally, situating cavities by a pair of rods or single rod defect in or adjacent to the waveguide, wave filtering characteristics of the cavities depending on the wavelength can be successfully confirmed.

As a useful numerical analysis technique, finite different time domain (FDTD) method [15] is powerful and widely applicable, for enabling to design various boundary shape of structure with multi-dimensional problems. In this paper, transmission characteristics of fork-shaped branch waveguide is numerically investigated by FDTD method [15]. In the first simulation, band-limited wave with time evolving envelope of sampling function is given as input. The transmitted frequency peaks were obtained by fast Fourier transform (FFT) In the second simulation, steady state electric field profile is sampled by giving a monochromatic signal with changing the wavelength range from 1,350 to 1,600 nm. The steady state wave is Fourier transformed and the peak of transmission spectra for given wavelength range is obtained. Both of the results show that obvious transmission peaks are observed in Fourier transformed wavelength domain. This suggests that the proposed waveguide structure have selectivity to the input signal wavelength, which can be applicable for wavelength filtering circuit component.

2 FDTD Method to Solve Two-Dimensional Maxwell's Equations

Maxwell's curl equations for electric field vector \mathbf{E} and magnetic field vector \mathbf{H} in lossless, isotropic, non-dispersive and non-conductive material are given as follows;

$$\nabla \times \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t}, \quad (1)$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \quad (2)$$

where ε is permittivity and μ is permeability of the space, respectively.

Assuming two dimensional uniform space along with z axis (i.e. $\partial/\partial z = 0$), Maxwell equations are decomposed into two sets of polarization. We analyze TE mode or E-wave which includes (H_x, E_y, H_z) as the component and propagates to x axis. For TE mode, Maxwell's equations are reduced to following sets;

$$\frac{\partial E_y}{\partial z} = \mu \frac{\partial H_x}{\partial t}, \quad (3)$$

$$\frac{\partial E_y}{\partial x} = -\mu \frac{\partial H_z}{\partial t}, \quad (4)$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = \varepsilon \frac{\partial E_y}{\partial t}. \quad (5)$$

For analyzing space and time, the electromagnetic field components and the medium profile are discretized and renewed step by step according th the leap-

frog algorithm [15] as followings;

$$H_x^{n+\frac{1}{2}}(i + \frac{1}{2}, k) = H_x^{n-\frac{1}{2}}(i + \frac{1}{2}, k) + \frac{\Delta t}{\mu \Delta z} \left\{ E_y^n(i + \frac{1}{2}, k + \frac{1}{2}) - E_y^n(i + \frac{1}{2}, k - \frac{1}{2}) \right\}, \quad (6)$$

$$H_z^{n+\frac{1}{2}}(i, k + \frac{1}{2}) = H_z^{n-\frac{1}{2}}(i, k + \frac{1}{2}) - \frac{\Delta t}{\mu \Delta x} \left\{ E_y^n(i + \frac{1}{2}, k + \frac{1}{2}) - E_y^n(i - \frac{1}{2}, k - \frac{1}{2}) \right\}, \quad (7)$$

$$E_y^{n+1}(i + \frac{1}{2}, k + \frac{1}{2}) = E_y^n(i + \frac{1}{2}, k + \frac{1}{2}) + \frac{\Delta t}{\varepsilon \Delta z} \left\{ H_x^{n+\frac{1}{2}}(i + \frac{1}{2}, k + 1) - H_x^{n+\frac{1}{2}}(i + \frac{1}{2}, k) \right\} - \frac{\Delta t}{\varepsilon \Delta x} \left\{ H_z^{n+\frac{1}{2}}(i + 1, k + \frac{1}{2}) - H_z^{n+\frac{1}{2}}(i, k + \frac{1}{2}) \right\}. \quad (8)$$

Time evolving electromagnetic field can be obtained by repeatedly solving above set of equation with any initial condition at beginning of simulation.

3 Two-Dimensional, Fork-Shaped Photonic Crystal Branch Waveguide with Silica Pillars

In Fig. 1, top view of periodic square lattice with a line-defect waveguide is shown. The longitudinal axis of the cylinder corresponds to polarization direction of electric field E_y of TE mode. Material of the cylinder is silica glass (SiO_2) with refractive index $n_1 = 3.6$ in background air $n_0 = 1.0$. The period $P = 551.8$ [nm], the radius $R = 0.2P$, discretized space grid $\Delta x = \Delta z = 9.85$ [nm], discretized time step $\Delta t = 0.021$ [fs] are used in the following simulation. Total number of pillars are 30×25 for horizontally and vertically in Fig. 1, which correspond to discrete grid numbers of $1,713 \times 1,433$ respectively.

In the simulation, input wavelength range is from 1,350 to 1,600 [nm], which correspond to range of operating wavelength in practical optical fiber communication system. For the lattice period P and the range of wavelength, the structure shows perfect photonic band gap, which means the optical wave can not penetrate into the periodic layer. Following the design, the incident wave from the input port can be guided along with line defect.

In Fig. 1, TE mode with components (H_x, E_y, H_z) is excited at port #1. The electric field E_y has Gaussian profile along x-axis as follows,

$$E_y(x, t) = E_0 \exp \left\{ - \left(\frac{x}{w_0} \right)^2 \right\} \sin \left(\frac{2\pi c_0 t}{\lambda} \right), \quad (9)$$

where, field amplitude $E_0 = 1.0$, half beam waist of Gaussian profile $w_0 = 275.8$ [nm], speed of light in vacuum $c_0 = 2.998 \times 10^8$ [m/s] and wavelength λ ,

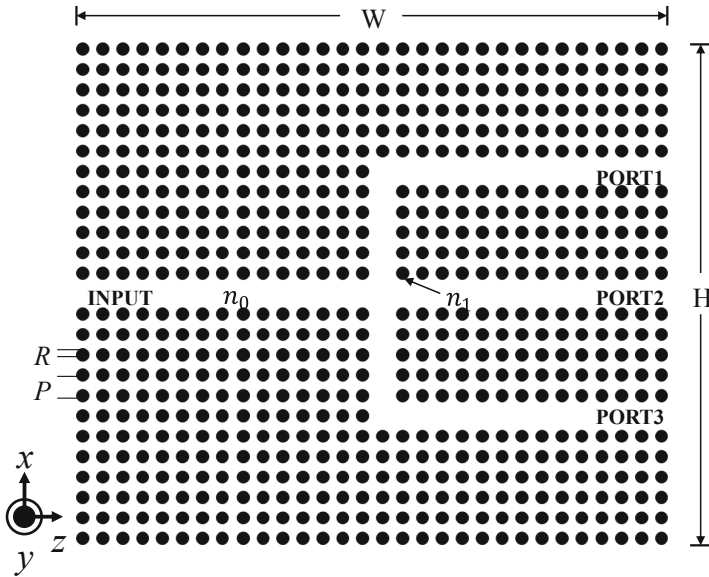


Fig. 1. Top view of square-lattice photonic crystal with fork-shaped line-defect branching waveguide.

respectively. Also shown in Fig. 1, three output ports are situated with fork-like shape. The input and output electric field E_y at each port are observed as function of time steps to be analyzed by fast Fourier transform technique.

4 Numerical Results

4.1 FFT Analysis of Output Electric Field in Fork-Shaped Branch Waveguide

In numerical analysis, the discrete time step is set to be $\Delta t = 5.0 \times 10^{-11}$ [s]. For band-limited spectrum with uniform and square profile where real part of the spectrum is unity and the imaginary part equals to zero, the time evolving input wave $f(t)$ is given by inverse Fourier transform of the spectrum as follows;

$$Re\{f(t)\} = -f_L \times Sa(2\pi f_L t) + f_U \times Sa(2\pi f_U t), \quad (10)$$

where

$$Sa(x) = \frac{\sin(x)}{x} \quad (11)$$

is a sampling function, f_L and f_U are lower and upper frequency [Hz] of the limited band, respectively. Here, $f_L = c_0/1,600$ [nm] and $f_H = c_0/1350$ [nm] are used for the flat square spectrum. The maximum input amplitude in the simulation comes at time $t = 100/f_C$ [s], where $1/f_C$ is time period for center frequency of the range and $f_C = (f_L + f_U)/2$.

From time-evolving data of FDTD analysis, sampled data are recorded every $\Delta t_{sample} = 50\Delta t = 2.5 \times 10^{-9}$ [s] with numbers of sample data $N_{sample} = 2^{16} = 65,536$. From these parameter, the frequency resolution $\Delta f = (\Delta t_{sample} \times N_{sample})^{-1} \simeq 1.9$ MHz.

4.2 Output Spectrum from Band-Limited Input Signal

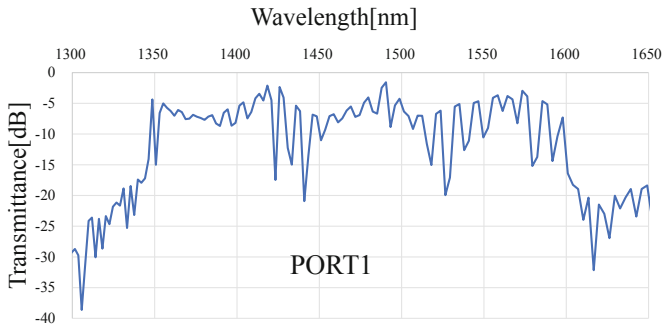
In Fig. 2(a) – (c), spectrum in each output port are plotted as function of wavelength. Spectrum in Figs. 2(a) and (c) are same as Port#1 and Port#3 are symmetrically situated. In Fig. 2(b) is output spectrum from Port#2, which locates at the center. In Fig. 2(b), four output typical wavelengths to show peaks exist at 1,356 [nm], 1,420 [nm], 1,490 [nm] and 1,570 [nm], respectively.

4.3 Output Spectrum from Steady State Profile of Monochromatic Continuous Input Signal

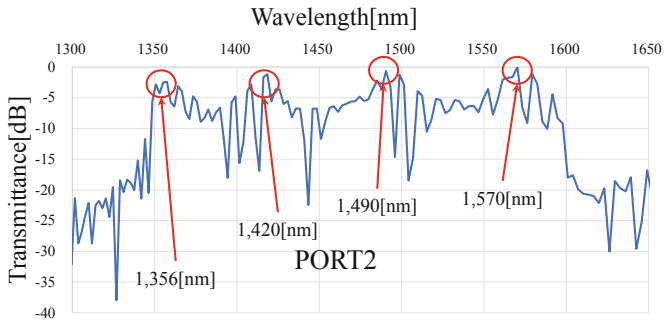
In this subsection, spectra which is obtained from steady state electric field profiles with monochromatic input signal is explained. In Fig. 3(a) – (c), transient input and output electric field profiles, an example of power spectrum for input signal of $\lambda = 1,350$ [nm], and the output spectra in wavelength range from 1,350 to 1,600 [nm], are depicted respectively. In Fig. 3(a), when input wavelength $\lambda = 1,350$ [nm], output field profile for continuous input electric field with $E_0 = 0.1$ is shown as a function of time step number. As is depicted, all output fields are converged after time step number of 20,000. Note that the output from PORT#1 and #3 are totally overlapping for symmetrical waveguide structure. In this figure, number of $2^{12} = 4,096$ data after 20,000 time step is Fourier transformed to obtain the spectrum. The power spectrum is shown in Fig. 3(b) as a function of wavelength, where unit of vertical axis is in arbitrary unit. It is obvious that extremely high peak exists at $\lambda = 1,350$ [nm]. Similarly, by changing input wavelength, output spectra from each output port is obtained. The power spectra as transmittance is plotted in Fig. 3(c). Please note again that the spectrum for PORT#1 and #3 are same as the output port locates with symmetry. We found that spectrum of PORT#2 which locates at center of three output ports has four local maximums.

In Table 1, the wavelength λ_{peak} to show local maximum output power is compared, and the error to result of steady state (Fig. 3(c)) is calculated. For 4 peak wavelengths, the error to Fig. 3(c) is less than 0.4%.

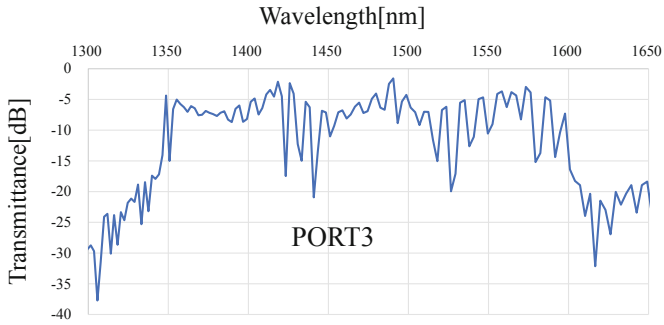
Let us consider about computation time to obtain these two simulation approaches, i.e., for case of band-limited input and of monochromatic continuous input. The former can obtain spectrum of entire wavelength range from 1,350 to 1,600 [nm] at once, while the latter requires 250 times longer computation of the former case. Though the latter case can get finer wavelength resolution of 1.0 [nm], for purpose of estimation of wavelength with output power peaks, we can get the results in much shorter time.



(a) Output spectrum in Port #1.

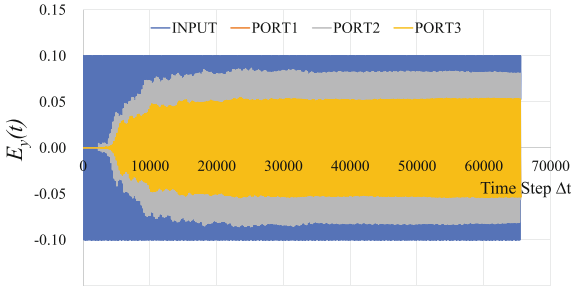


(b) Output spectrum in Port #2.

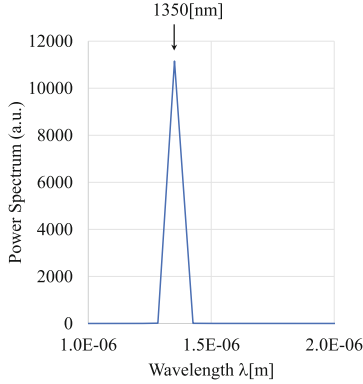


(c) Output spectrum in Port #3.

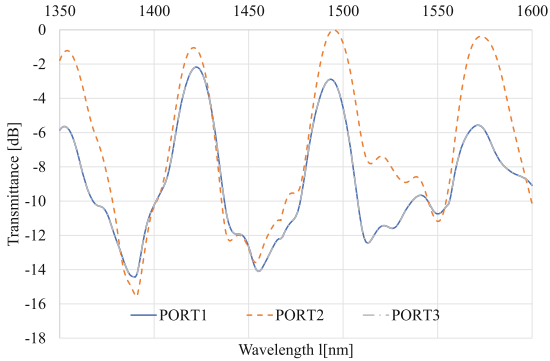
Fig. 2. Output spectrum in (a) Port #1 , (b)Port #2, and (c) Port #3, respectively.



(a) Electric field profiles as a function of time step number.



(b) Example of output power spectrum at PORT #1 (center) for input wavelength $\lambda = 1350\text{nm}$.



(c) Transmittance of proposed fork-shaped waveguide as a function of input wavelength.

Fig. 3. Numerical results for continuous monochromatic input wavelength. (a) Electric field profile of input and each output port, (b) example of Fourier transformed power spectrum, and (c) power spectrum as transmittance as function of input wavelength in range of 1,350–1600 nm, respectively.

Table 1. Comparison of λ_{peak} in PORT#2 from band-limited input and monochromatic continuous input. The error is estimated to λ_{peak} in Fig. 3(c).

λ_{peak} [nm] in Fig. 2(b)	1,355	1,422	1,495	1,574
λ_{peak} [nm] in Fig. 3(c)	1,360	1,420	1,490	1,570
Error in %	0.37	0.14	0.33	0.25

5 Conclusion and Future Subject

Output spectrum of fork-shaped branch waveguide in square lattice pillar type photonic crystal waveguide are numerically demonstrated. The time evolving electric field profiles and output spectra in two symmetrical ports coincided very well. At the center output port, the spectra showed four typical peaks both for case of band-limited input and of monochromatic continuous input. As monochromatic input can obtain output power peaks in finer wavelength resolution with much longer computation time, the band-limited input can estimate the output power peaks with good accuracy and in quite shorter computation time compared with former case.

Acknowledgment. Author expresses my appreciation to Mr. H. Ikeda, Mr. T. Haradaguchi, Mr. T. Fujiki, and Mr K. Mutoh of Fukuoka Institute of Technology as part of their undergraduate research under supervision by the author in 2021–22.

References

1. Yasumoto, K. (ed.): Electromagnetic Theory and Applications for Photonic Crystals. CRC PRESS, Boca Raton (2006)
2. Inoue, K., Ohtaka, K. (eds.): Photonic Crystals - Physics, Fabrication and Applications. Springer, New York (2004)
3. Noda, S., Baba, T. (eds.): Roadmap on Photonic Crystals. Kluwer Academic Publishers, Alphen aan den Rijn (2003)
4. Joannopoulos, J.D., Meade, R.D., Winn, J.N.: Photonic Crystals. Princeton University Press, New Jersey (1995)
5. Maeda, H., Haari, K., Meng, X.Z., Higashinaka, N.: Signal routing by dispersive medium. In: Barolli, L., Xhafa, F., Conesa, J. (eds.) BWCCA 2017. LNDECT, vol. 12, pp. 764–773. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-69811-3_68
6. Maeda, H., Meng, X.Z., Haari, K., Higashinaka, N.: Signal routing by cavities in photonic crystal waveguide. In: Barolli, L., Xhafa, F., Javaid, N., Spaho, E., Kolicic, V. (eds.) EIDWT 2018. LNDECT, vol. 17, pp. 765–772. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75928-9_68
7. Haari, K., Maeda, H., Meng, X.Z., Higashinaka, N.: Numerical analysis of optical duplexer composed of photonic crystal with square lattice. In: Barolli, L., Leu, F.-Y., Enokido, T., Chen, H.-C. (eds.) BWCCA 2018. LNDECT, vol. 25, pp. 548–558. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02613-4_49

8. Haari, K., Higashinaka, N., Meng, X.Z., Maeda, H.: Numerical analysis of optical duplexer composed of dispersive and nonlinear dielectric in two-dimensional photonic crystal waveguide with square lattice. In: Barolli, L., Takizawa, M., Xhafa, F., Enokido, T. (eds.) WAINA 2019. AISC, vol. 927, pp. 275–285. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15035-8_26
9. Haari, K., Higashinaka, N., Maeda, H., Meng, X.: Numerical analysis of optical wavelength switching by dispersive and nonlinear dielectric in 2D photonic crystal. In: Proceedings of the 8th Asia-Pacific Conference on Antennas and Propagation (APCAP 2019), pp.152-153 (2019)
10. Maeda, H., Higashinaka, N., Ochi, A.: Numerical analysis of Fano resonator in 2D periodic structure for integrated microwave circuit. In: Barolli, L., Nishino, H., Enokido, T., Takizawa, M. (eds.) NBiS - 2019 2019. AISC, vol. 1036, pp. 630–637. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-29029-0_62
11. Higashinaka, N., Maeda, H.: Routing of optical baseband signal depending on wavelength in periodic structure. In: Barolli, L., Hellinckx, P., Enokido, T. (eds.) BWCCA 2019. LNNS, vol. 97, pp. 621–629. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-33506-9_56
12. Higashinaka, N., Maeda, H.: Input amplitude dependency of duplexer with dispersive and nonlinear dielectric in 2-D photonic crystal waveguide. In: Barolli, L., Okada, Y., Amato, F. (eds.) EIDWT 2020. LNDECT, vol. 47, pp. 226–236. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39746-3_25
13. Maeda, H., Higashinaka, N.: Wavelength tuning of output optical signal through resonant filter for WDM system by periodic structure composed of silica glass. In: Barolli, L., Li, K.F., Enokido, T., Takizawa, M. (eds.) NBiS 2020. AISC, vol. 1264, pp. 488–497. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-57811-4_49
14. Maeda, H.: Numerical analysis of photonic crystal waveguide with fork-shaped branch. In: Proceedings of NBiS-2022, accepted and to be presented (2022)
15. Taflove, A.: Advances in Computational Electrodynamics - The Finite-Difference Time-Domain Method. Artech House, Norwood (1995)



Design and Implementation of a Platform for MOAP Robots

Keita Matsuo^{1(✉)}, Elis Kulla², and Leonard Barolli¹

¹ Department of Information and Communication Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
`{kt-matsuo,barolli}@fit.ac.jp`

² Department of System Management, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
`kulla@fit.ac.jp`

Abstract. Recently, various communication technologies have been developed in order to satisfy the requirements of many users. Especially, mobile communication technology continues to develop rapidly and Wireless Mesh Networks (WMNs) are attracting attention from many researchers in order to provide cost efficient broadband wireless connectivity. The main issue of WMNs is to improve network connectivity and stability in terms of user coverage. In our previous work, we presented Moving Omnidirectional Access Point (MOAP) robots, which can move omnidirectionally in the real space to provide good communication and stability for WMNs. In this paper, we present the design and implementation of a platform for controlling MOAP robots. The implemented platform is able to calculate the optimal number of MOAP robots and their positions considering clients positions and communication area. The implemented platform can get accurate robot position by using supersonic sensors and move MOAP robot to appropriate positions to provide good communication. The experimental results show that the error distance of both X and Y directions is within 1 [cm].

1 Introduction

Recently, communication technologies have been developed in order to satisfy the requirements of many users. Especially, mobile communication technologies continue to develop rapidly and has facilitated the use of laptops, tablets and smart phones in public spaces [5]. In addition, Wireless Mesh Networks (WMNs) [2] are becoming an important network infrastructure. These networks are made up of wireless nodes organized in a mesh topology, where mesh routers are interconnected by wireless links and provide Internet connectivity to mesh clients.

WMNs provide cost efficient broadband wireless connectivity. The main issue of WMNs is to improve network connectivity and stability in terms of user coverage. This problem is very closely related to the family of node placement problems in WMNs [6, 10, 12]. In these papers it is assumed that routers move by themselves or by using network simulator moving models.

In our research, we consider a moving robot as network device. In order to realize a moving access point, we implemented a Moving Omnidirectional Access Point robot (called MOAP robot). It is important that the MOAP robot moves to an accurate position in order to have a good connectivity. Thus, the MOAP robot can provide good communication and stability for WMNs. Moreover, we need to set minimum Transmission (TX) power for MOAP robot in order to avoid interference and save energy.

In this paper, we present the design and implementation of a platform for controlling MOAP robots. The platform is able to calculate the optimal number of MOAP robots and their positions considering clients positions and communication area. The implemented platform can get accurate robot position by using supersonic sensors and move MOAP robots to appropriate positions to provide good communication.

The rest of this paper is structured as follows. In Sect. 2, we introduce the related work. In Sect. 3, we propose a platform for MOAP robots. In Sect. 4, we describe the experimental environment and results. Finally, conclusions and future work are given in Sect. 5.

2 Related Work

Different techniques are developed to solve the problem of moving robots position. One of important research area is indoor position detection, because the outdoor position can be detected easily by using GPS (Global Positioning System). However, in the case of indoor environment, we can not use GPS. So, it is difficult to find the target position.

Asahara et al. [3] proposed to improve the accuracy of the self position estimation of a mobile robot. A robot measures the distance to an object in the mobile environment by using a range sensor. Then, the self position estimation unit estimates the position of the mobile robot based on the selected map data and range data obtained by the range sensor. Wang et al. [13] proposed the ROS (Robot Operating System) platform. They designed a WiFi indoor initialize positioning system by triangulation algorithm. The test results show that the WiFi indoor initialize position system combined with AMCL (Adaptive Monte Carlo Localization) algorithm can be accurately positioned and has high commercial value.

Nguyen et al. [11] proposed a low speed vehicle localization using WiFi fingerprinting. In general, these researches rely on GPS in fusion with other sensors to track vehicle in outdoor environment. However, as indoor environment such as car park is also an important scenario for vehicle navigation, the lack of GPS poses a serious problem. For this reason, the authors used an ensemble classification method together with a motion model in order to deal with the issue. Experimental results show that proposed method is capable of imitating GPS behavior on vehicle tracking.

Ban et al. [4] proposed indoor positioning method integrating pedestrian Dead Reckoning with magnetic field and WiFi fingerprints. Their proposed

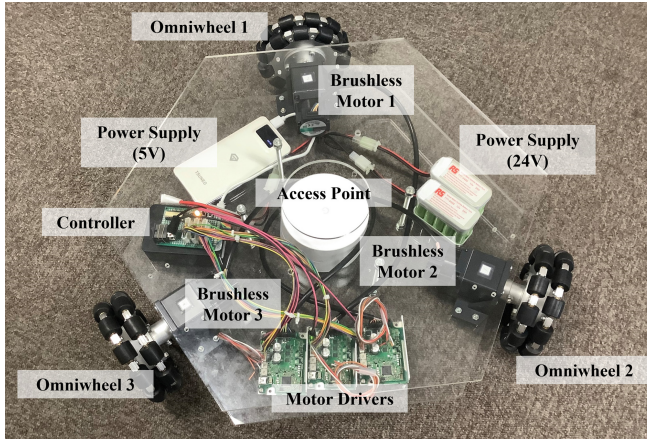


Fig. 1. Implemented MOAP robot.

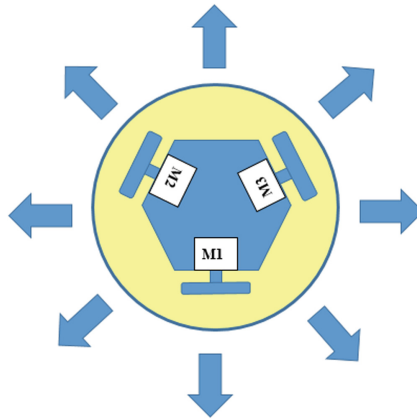


Fig. 2. Movement of our implemented MOAP robot.

method needs WiFi and magnetic field fingerprints, which are created by measuring in advance the WiFi radio waves and the magnetic field in the target map. The proposed method estimates positions by comparing the pedestrian sensor and fingerprint values using particle filters. Matsuo et al. [7, 8] implemented and evaluated a small size omnidirectional wheelchair. Also, Matsuo et al. [9] evaluated elbow and silhouette theories to decide MOAP robots positions to have better communication environment.

3 Proposed Platform for MOAP Robots

In this section, we describe the implemented MOAP robot platform. We show the implemented MOAP robot in Fig. 1. The MOAP robot can move omnidi-

Table 1. Specifications of MOAP robot.

Items	Specifications
Length	490.0 [mm]
Width	530.0 [mm]
Height	125.0 [mm]
Brushless Motor	BLHM015K-50 (Orientalmotor corporation)
Motor Driver	BLH2D15-KD (Orientalmotor corporation)
Controller	Raspberry Pi 3 Model B+
Power Supply	DC24V Battery
PWM Driver	Pigpio (The driver can generate PWM signal with 32 line)

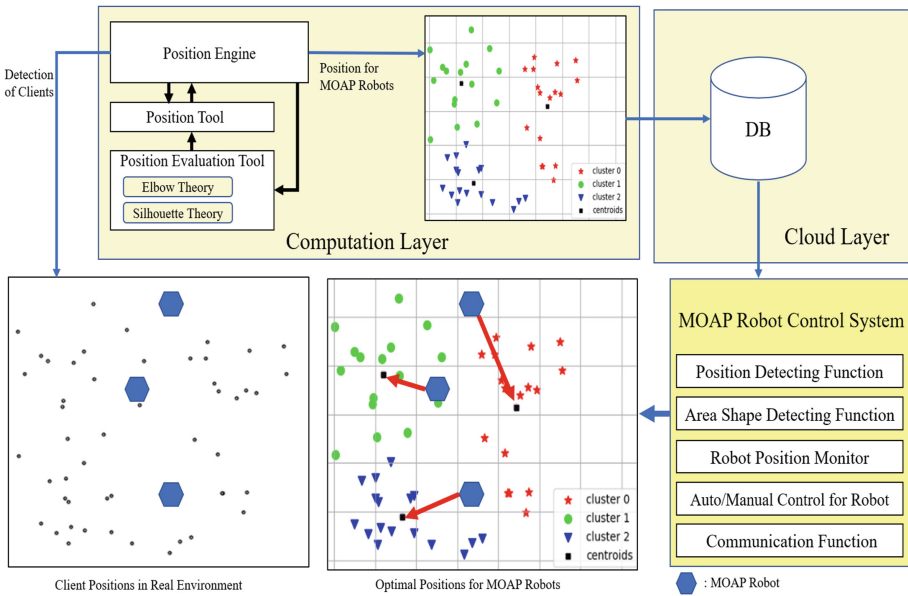


Fig. 3. Proposed platform.

rectionally keeping the same direction and can provide access points for network devices. In order to realize our proposed MOAP robot, we used omniwheels which can rotate omnidirectionally in front, back, left and right. The movement of the MOAP robot is shown in Fig. 2. We would like to control the MOAP robot to move accurately in order to offer a good environment for communication.

Our implemented MOAP robot has 3 omniwheels, 3 brushless motors, 3 motor drivers and a controller. The MOAP robot requires 24V battery to move and 5V battery for the controller. We show the specifications of MOAP robot in Table 1.

We show the proposed platform in Fig. 3, which can calculate the optimal number of MOAP robots and their position. The platform offers users a good communication environment by moving MOAP robots to appropriate positions.

First, the platform detects the clients. Next it uses K-means clustering by the position tool. Then, it calculates the optimal MOAP robot positions by using Elbow and Silhouette theories. After deciding MOAP robot positions, the platform uploads the position data of MOAP robots to the Data Base (DB). Finally, the platform uses the position data to move the MOAP robots to better positions for keeping a good communication environment.

In order to decide optimal positions and the number of MOAP robot, the platform considers Elbow and Silhouette theories. Elbow and Silhouette theories use K-means clustering. We show K-means function in Eq. (1). In this case, C_i means i th cluster and x_{ij} is j th of i th data. K is the number of clusters. Ideal clustering is achieved when the value of Eq. (1) will be minimized.

$$\min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\} \quad (1)$$

We show the K-means clustering in Fig. 4. The dots show clients. In Fig. 4 (a), we deployed 150 clients in random way on 2D space (100 m × 100 m). After that, we used K-means clustering as shown in Fig. 4 (b). We consider that the centroids can communicate with each-other for this scenario.

3.1 Elbow Theory

Elbow theory uses the distance between centroid and clients (see Fig. 4 (b)). In Eq. (2), *All_Distance* means total distance between each centroid and clients in the cluster. If there is only one cluster, the *All_Distance* value is maximum. When the cluster number increases the *All_Distance* value is decreased. The relation between *All_Distance* and the number of clusters is shown in Fig. 5. From this figure, we can see that the optimal number of clusters is 3. This is the elbow value.

$$All_Distance = \sum_{k=1}^K \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \quad (2)$$

3.2 Silhouette Theory

Silhouette theory uses the value of Silhouette coefficient. We show Silhouette coefficients in Eq. (3), where $a^{(i)}$ is the average distance between i -th client and other clients in the same cluster and $b^{(i)}$ is the average distance between i -th client and other clients in the nearest cluster. The $s^{(i)}$ shows the degree of success or failure of clustering. The value range for $s^{(i)}$ is -1 to 1 . If $s^{(i)}$ value is near to 1 the clients in the same cluster are very close to each other. When $s^{(i)}$ value is

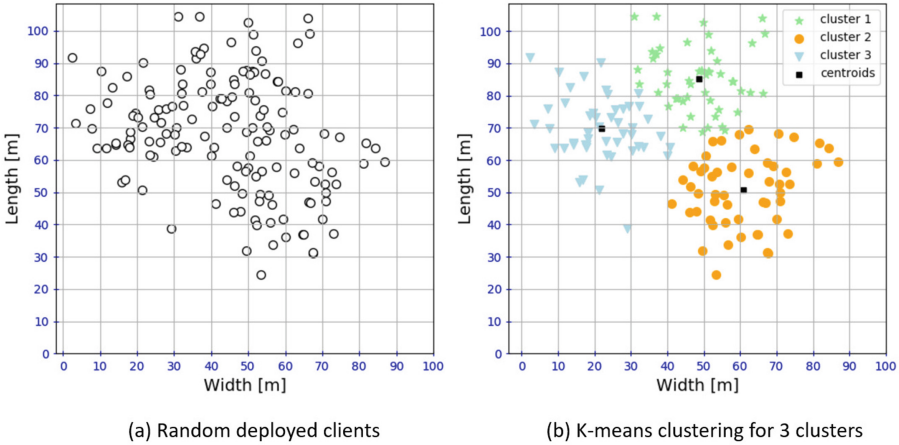


Fig. 4. Simulation with 3 clusters.

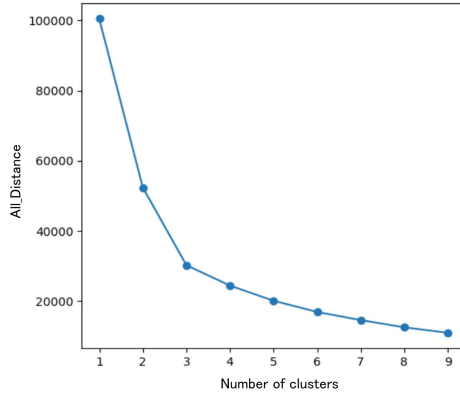


Fig. 5. Relation between All_Distance and number of clusters (for 3 clusters).

0, the clients are located on the border between clusters. Also, when the value of $s^{(i)}$ is negative, the client does not belong to an appropriate cluster.

We show the K-means clustering and Silhouette coefficients with 8 clusters in Fig. 6 and Fig. 7. In Fig. 6 (a), we deployed 800 clients on the 2D space (150m × 120m) randomly. After that we clustered the clients by 8 clusters using K-means clustering in order to analyze the clusters using the Silhouette theory (see Fig. 6 (b)).

In Fig. 7 are shown silhouette coefficients of 8 clusters, where the vertical line means average value. The thickness of each clusters shows the number of clients in the cluster. The silhouette coefficient is high when the thickness are almost the same.

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{a^{(i)}, b^{(i)}\}} \tag{3}$$

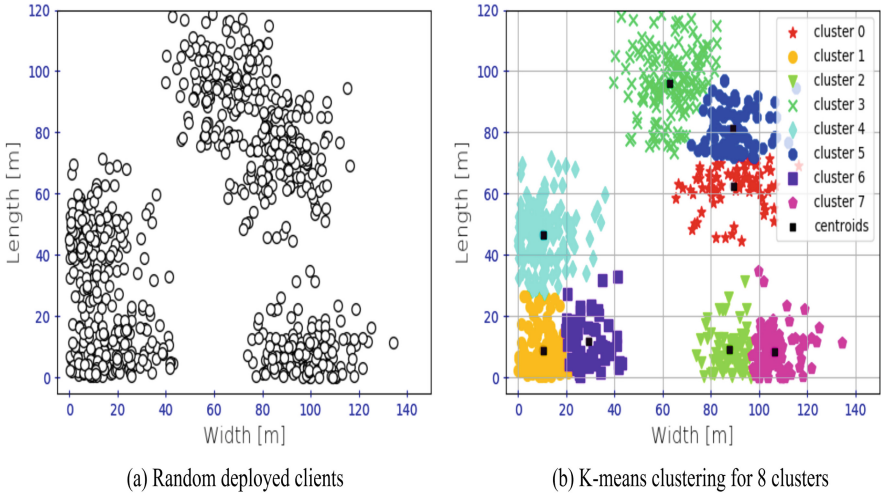


Fig. 6. Simulation with 8 clusters.

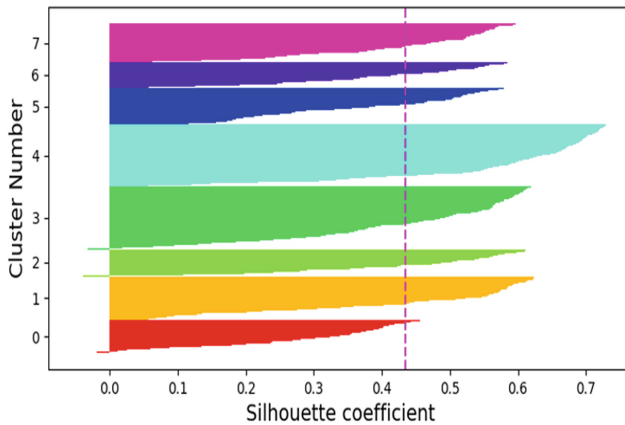


Fig. 7. Silhouette coefficient with 8 clusters.

3.3 Implementation of Proposed Platform for Controlling MOAP Robots

In Fig. 8 is shown the implemented platform for controlling MOAP robots considering Elbow and Silhouette theories. The red and yellow square marks show the location of two sensors attached to MOAP robot. The mid-point of the line between red and yellow square is considered as the robot position.

In Fig. 9 is shown position detecting system and real supersonic sensors made by Marvelmind [1]. Also, there are stationary beacons and mobile beacons in Fig. 9. In Table 2 are shown the specifications of supersonic sensors. The control

Table 2. Specifications of supersonic sensors.

Items	Specifications
Distance between beacons	Up to 50 [m]
Coverage area	Up to 1000 [m ²] with 4 beacons
Location update rate	0.05 – 25 [Hz]
Power supply	LiPol battery 1000 [mAh] (Internal)
Beacon size	W55 × L55 × H33 [mm] (with antenna: H65 [mm])
Weight	59 [g] (including battery)

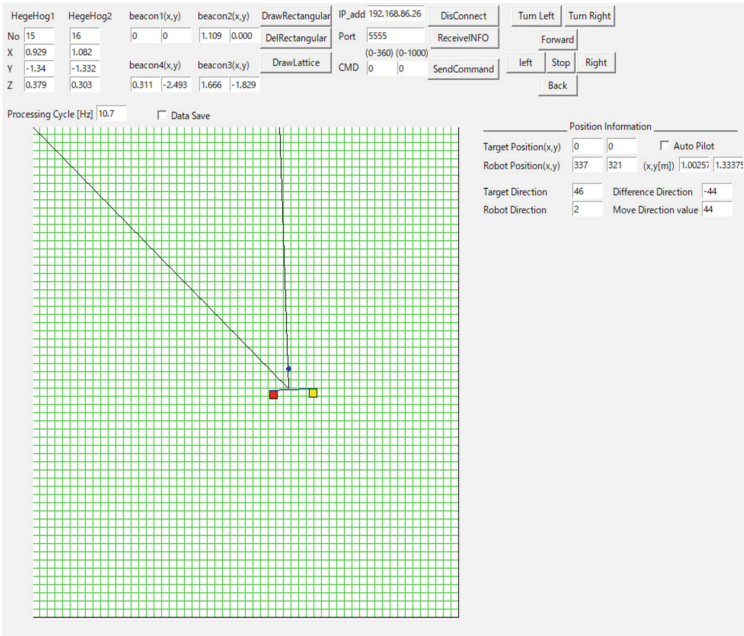


Fig. 8. Implemented platform for controlling MOAP robot.

system can detect different shapes of communication area and move the MOAP robot in a good position to cover the clients as shown in Fig. 10.

4 Experimental Results

To confirm the accuracy of the MOAP robot position, we carried out experiments to detect the MOAP robot positions as shown in Fig. 11. The MOAP robot moved from detected point 1 to point 4 and we measured 200 times each position at 4 points. The numbers in parentheses are for each coordinates (x, y). The top left side coordinates are (0, 0) and is considered as origin.

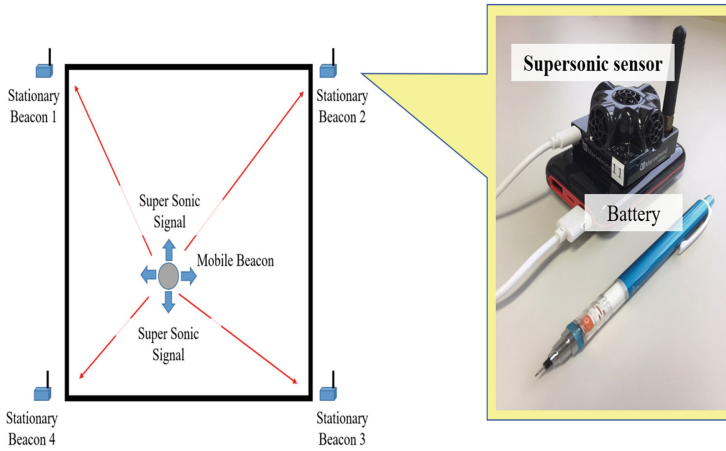
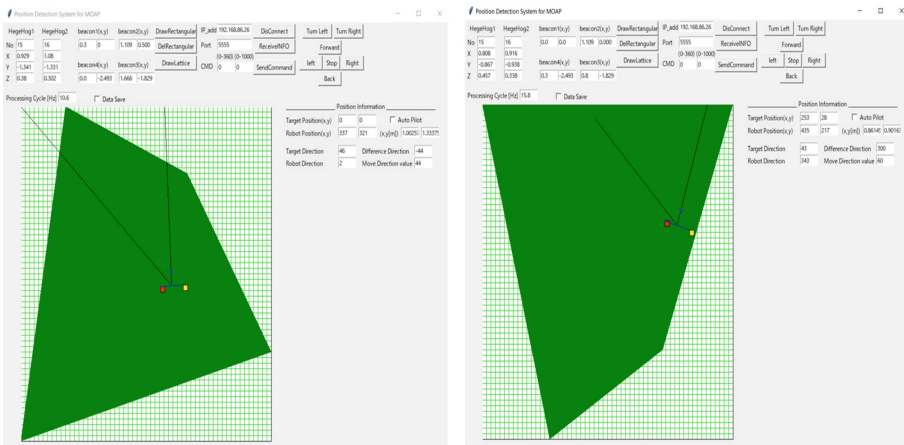


Fig. 9. Implemented MOAP robot position detecting system.



(a)

(b)

Fig. 10. Example of detecting different shapes of communication area.

We show the experimental results in Fig. 12. In Fig. 12(a) is shown the distance of x direction and in Fig. 12(b) is shown the distance of y direction from the origin. We evaluated the error distance on both X axis and Y axis. The average error distance on X axis is 0.007 [m] while for Y axis is 0.002 [m].

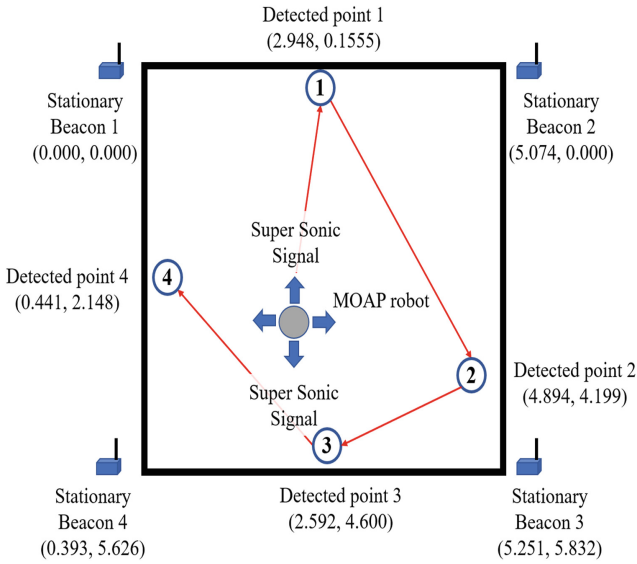


Fig. 11. Experiment for position detection.

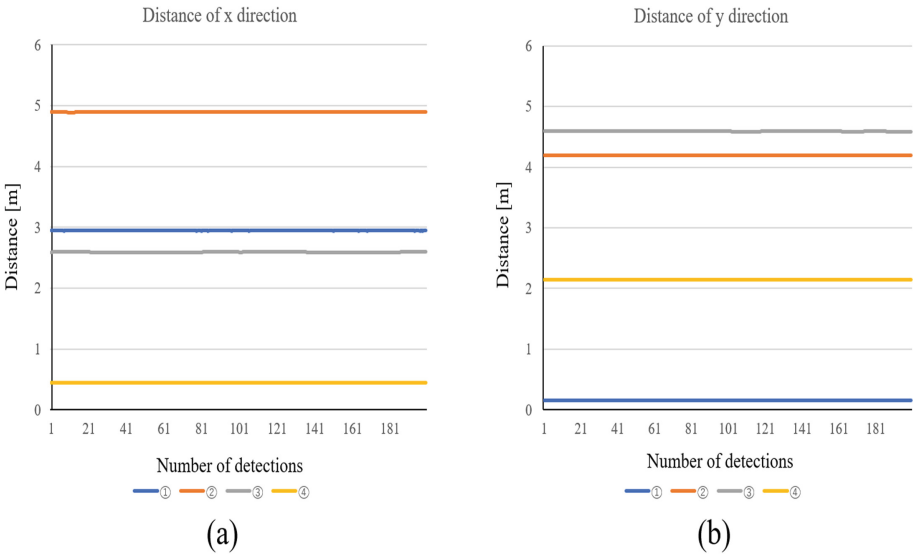


Fig. 12. Experimental results of detecting MOAP robot positions.

5 Conclusions and Future Work

In this paper, we introduced our implemented platform for controlling MOAP robots. We showed some of the previous works and discussed the related problems

and issues. Then, we presented some feature of MOAP robot and proposed Elbow and Silhouette theories to calculate the optimal number of MOAP robots in order to have a good communication environment for WMNs.

The proposed platform could find the optimal number of MOAP robots positions and has a good accuracy to control MOAP robots. The experimental results show the implemented platform can get accurate robot position by using super-sonic sensors and provides a good communication environment. Also, the error distance of both X and Y direction were within 1 [cm].

In the future work, we would like to add other functions to the platform and carry out extensive simulations and experiments to evaluate the proposed platform.

References

1. Marvelmind. <https://marvelmind.com/>
2. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. *Comput. Netw.* **47**(4), 445–487 (2005)
3. Asahara, Y., Mima, K., Yabushita, H.: Autonomous mobile robot, self position estimation method, environmental map generation method, environmental map generation apparatus, and data structure for environmental map, 19 January 2016, uS Patent 9,239,580
4. Ban, R., Kaji, K., Hiroi, K., Kawaguchi, N.: Indoor positioning method integrating pedestrian dead reckoning with magnetic field and WiFi fingerprints. In: 2015 Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU), pp. 167–172 (2015)
5. Hamamoto, R., Takano, C., Obata, H., Ishida, K., Murase, T.: An access point selection mechanism based on cooperation of access points and users movement. In: 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 926–929 (2015)
6. Maolin, T.: Gateways placement in backbone wireless mesh networks. *Int. J. Commun. Netw. Syst. Sci.* **2**(01), 44–50 (2009)
7. Matsuo, K., Barolli, L.: Design and implementation of an omnidirectional wheelchair: control system and its applications. In: Proceedings of the 9th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA-2014), pp. 532–535 (2014)
8. Matsuo, K., Liu, Y., Elmazi, D., Barolli, L., Uchida, K.: Implementation and evaluation of a small size omnidirectional wheelchair. In: Proceedings of the IEEE 29th International Conference on Advanced Information Networking and Applications Workshops (WAINA-2015), pp. 49–53 (2015)
9. Matsuo, K., Mitsugi, K., Toyama, A., Kulla, E., Barolli, L.: A simulation system for optimal positions of MOAP robots using elbow and silhouette theories: simulation results considering minimum transmission power of MOAP robots. In: Barolli, L. (ed.) BWCCA 2021. LNNS, vol. 346, pp. 321–332. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-90072-4_35
10. Muthaiah, S.N., Rosenberg, C.: Single gateway placement in wireless mesh networks. *Proc. ISCN* **8**, 4754–4759 (2008)
11. Nguyen, D., Recalde, M.E.V., Nashashibi, F.: Low speed vehicle localization using WiFi fingerprinting. In: 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 1–5 (2016)

12. Oda, T., Barolli, A., Spaho, E., Xhafa, F., Barolli, L., Takizawa, M.: Performance evaluation of wmn using wmn-ga system for different mutation operators. In: 2011 14th International Conference on Network-Based Information Systems. pp. 400–406 (Sep 2011)
13. Oda, T., Barolli, A., Spaho, E., Xhafa, F., Barolli, L., Takizawa, M.: Performance evaluation of WMN using WMN-GA system for different mutation operators. In: 2011 14th International Conference on Network-Based Information Systems, pp. 400–406 (2011)



Design of an Intelligent Robotic Vision System for Optimization of Robot Arm Movement

Chihiro Yukawa¹, Nobuki Saito¹, Aoto Hirata¹, Kyohei Toyoshima¹, Yuki Nagai¹,
Tetsuya Oda²(✉), and Leonard Barolli³

¹ Graduate school of Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan

{t22jm19st, t21jm01md, t21jm02zr, t22jm24jd, t22jm23rv}@ous.jp

² Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama-shi 700-0005, Japan
oda@ice.ous.ac.jp

³ Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan
barolli@fit.ac.jp

Abstract. The goal of Industry 4.0 is to achieve a higher level of operational efficiency and productivity, as well as a higher level of automatization. The automation is considered in the manufacturing industry to improve the efficiency of production processes. Also, the measurement at the nano-level on the surface of the target object by a machine has been considered for automation, but there are problems such as the need for high cost and a large amount of time for measurement. In this paper, we propose a robot vision system based on an intelligent algorithm for recognizing micro-roughness on arbitrary surfaces. The proposed system is inexpensive, make quick measurement and is capable of autonomously recognizing micro-roughness to improve the efficiency of production processes. The experimental results show that the Hill Climbing (HC) algorithm can reduce the movement vibration.

1 Introduction

The automation is considered by Industry 4.0 [1] in the manufacturing industry to improve the efficiency of production processes. Also, the measurement of micro-roughness on the target surface by machines is being considered for automation [2–8]. However, there are problems such as the high cost and a lot of time required for measurement. In addition, most of the processing skills related to micro-roughness and convexities have been performed manually by craftsmen who have many years of experience.

It is important to transmit the skills to the next generation, but it takes a lot of time and experience to learn the skills as artisans, therefore continuous employment is required from the young generation. But there are few people who want to become artisans, and it is difficult to hire young people, resulting in a chronic labor shortage. On the other hand, the skills of artisans can be technically reproduced by applying current

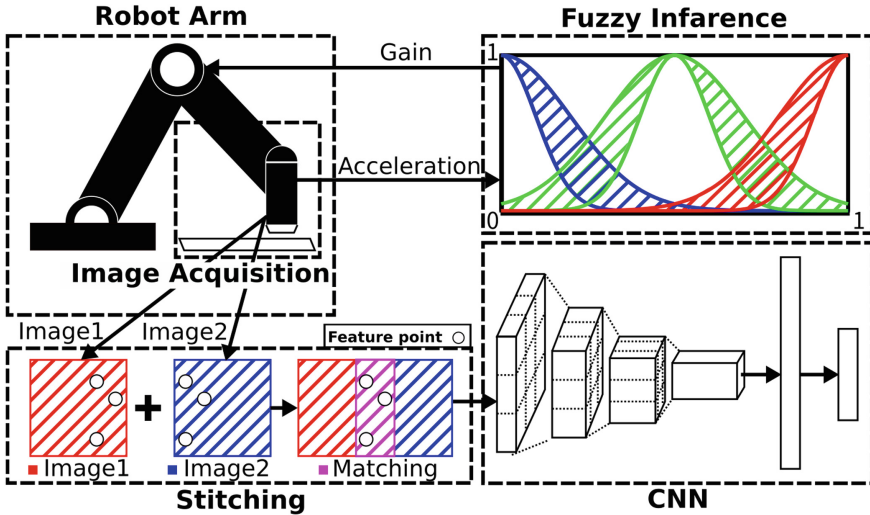


Fig. 1. Proposed system.

machine learning and control technologies. The robotization will become possible by imitating the skills of artisans by intelligent algorithms [9–12].

In this paper, we present a robot vision system based on an intelligent algorithm for recognizing micro-roughness on arbitrary target surfaces. The proposed system is inexpensive, make quick measurement and is capable of autonomously recognizing micro-roughness to improve the efficiency of production processes. The experimental results show that the Hill Climbing (HC) algorithm can reduce the movement vibration.

The structure of the paper is as follows. In Sect. 2, we present the proposed system. In Sect. 3, we present the optimization of robot arm movement. In Sect. 4, we discuss the experimental results. Finally, conclusions and future work are given in Sect. 5.

2 Proposed System

The robotic vision system for recognizing micro-roughness on the arbitrary target surface is shown in Fig. 1. The proposed system considers images from various angles with the 4° of freedom. For the robot arm, we propose a vibration suppression method based on fuzzy inference [13–17] to improve the recognition rate of micro-roughness. An electron microscope is mounted to the edge of the robot arm. The images received by the electron microscope are reconstructed by stitching. Then, the stitched images are sent to a Deep Learning Network (Conventional Neural Network: CNN) for object recognition. We use Hill Climbing (HC) for the reduction of the movement vibration of robot arm.

2.1 Servo Motors Vibration Reduction Method

The robot arm has vibration and low accuracy of movement [18]. But, by using the training datasets for image recognition, it is possible to improve the recognition rate by suppressing the movement vibration of the robot arm [19]. Therefore, we propose a method for reduction of vibration of the robot arm considering sensing, fuzzy inference, and servomotor control to improve the recognition rate of micro-roughness. The sensing module sends to the Jetson the acceleration values X , Y and Z -axis from the accelerometer GY-521 mounted to the edge of the robot arm. Also, it sends the sensing data to controller via serial transmission.

In the fuzzy inference are inserted the values of error and angle. The error is the average squared error of the X , Y and Z -axis, where the true value is the average value of the acceleration received from the accelerometer. We use Interval Type-2 Fuzzy Sets (IT2FS) [20, 21, 26], and the output is determined by the Enhanced Iterative Algorithm with Stop Condition (EIASC) method [24, 25]. The IT2FS output is the gain of servo motor. The input membership functions are the angle of robot arm and the error of acceleration. The gain affects the responsiveness of the servo mechanism. The servo motor controls the angle based on the gain of fuzzy inference output. In the vibration suppression method, the gain is gradually decreased from the starting position to the specified angle to suppress the vibration. The gain of the robot arm is gradually reduced from the start position.

2.2 Image Stitching

In image stitching, the feature points are detected in multiple images by the electron microscope and matching is performed based on the feature points. It is possible to obtain a single high-resolution image of the entire acquired object by stitching each image and the duplicate recognized images should be prevented [25]. In addition, the prediction of object detection time should be reduced.

2.3 Object Detection

The proposed system recognizes micro-roughness on the surface by YOLOv5. We use a deep learning model to perform object detection and class classification. A large number of images containing micro-roughness are required to train the model. However, since there are only a few factories and other production sites to process micro-roughness only few images can be used for model training [24]. Therefore, we created a dataset of images containing micro-roughness for training YOLOv5. In addition, we consider transfer learning [22, 23], which is a method of transferring a model that has already been trained for another problem to improve the recognition rate of micro-roughness and reduce the training time.

3 Optimization of Robot Arm Movement

In this section, we consider the optimization of robot arm movement for vibration reduction. The pseudo-code of the proposed method is shown in Algorithm 1. The proposed method determines the route of robot arm movement by using image acquisition points. The HC algorithm is used for vibration reduction. First, the route of robot arm movement is determined randomly in all image acquisition points. Next, two points replace randomly selected points in order to set a new route. Also, vibration is calculated for the robot arm motion in a new route. The vibration is added to each acceleration values in X , Y and Z -axis from the accelerometer. If the acceleration of the new route reduces the acceleration of the previous route, the route is updated to the new route. These steps are repeated for a set number of times.

Algorithm 1. Optimization Route Order for Robot Arm

```

1: Generate initial solution  $R$ .  $R$  is route of robot arm.
2: for Motion robot arm in  $R$  do
3:    $A \leftarrow A + \text{Acceleration}$ 
4: for  $i = 0$  to 100 do
5:    $j, k \leftarrow$  Randomly number of 0 to route length.
6:    $R' \leftarrow$  Swap  $R[j], R[k]$ 
7:   for Motion robot arm in  $R'$  do
8:      $A' \leftarrow A' + \text{Acceleration}$ 
9:     if  $A' < A$  then
10:       $R, A \leftarrow R', A'$ 
11: return Final solution  $R$ .

```

4 Experimental Results

The experimental environment is shown in Fig. 2. Experiments are performed on metal surfaces. As the robot arm is used uArm Swift Pro Standard with 4° of Freedom and as the microscope is used 5MP Digital Microscope USB 2.0. The experimental results for the optimization of the robot arm movement are shown in Fig. 3 and the visualization results are shown in Fig. 4. For the experiment, there are 9 image acquisition points and the number of iterations is 100. From the experimental results, we conclude that the HC algorithm can reduce the movement vibration.

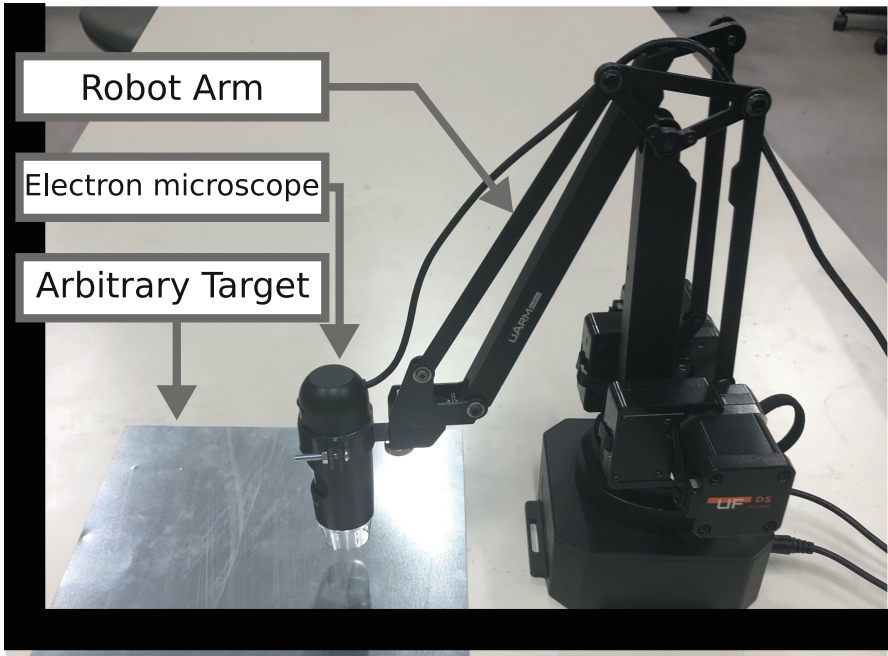


Fig. 2. Experimental environment.

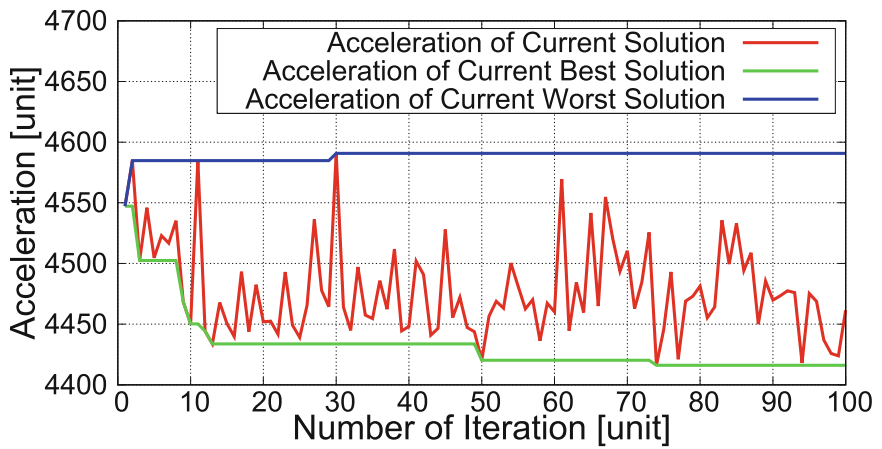


Fig. 3. Experimental results.

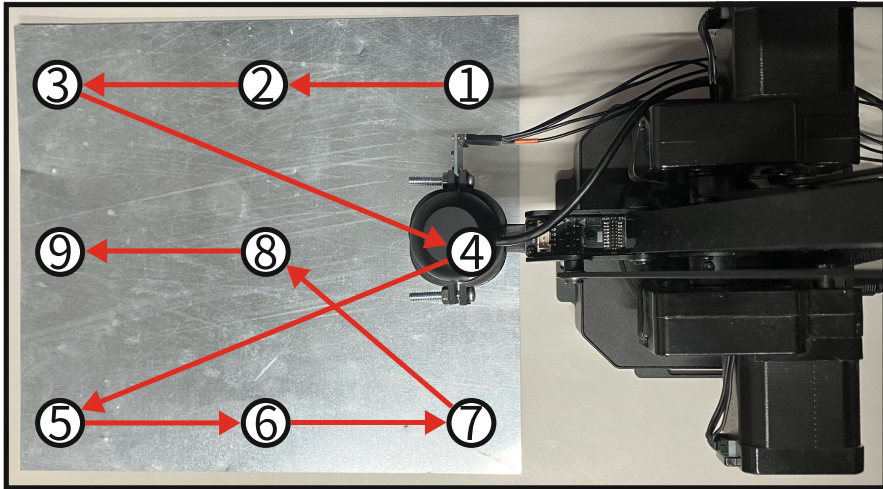


Fig. 4. Visualization results.

5 Conclusions

In this paper, we proposed a robot vision system based on an intelligent algorithm for recognizing micro-roughness on arbitrary surfaces. The proposed system is inexpensive, make quick measurement and is capable of autonomously recognizing micro-roughness to improve the efficiency of production processes. The experimental results show that the HC algorithm can reduce the movement vibration. In the future, we would like to increase the number of image acquisition points.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 20K19793.

References

1. Dalenogare, L., et al.: The expected contribution of industry 4.0 : *Int. J. Prod. Econ. (IJPE-2018)* **204**, 383-394 (2018)
2. Shang, L., et al.: Detection of rail surface defects based on CNN image recognition and classification. In: *The IEEE 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 45-51 (2018)
3. Li, J., et al.: Real-time detection of steel strip surface defects based on improved yolo detection network. *IFAC-PapersOnLine* **51**(21), 76–81 (2018)
4. Oda, T., et al.: Design and implementation of a simulation system based on deep Q-network for mobile actor node control in wireless sensor and actor networks: In: *Proceedings of The IEEE 31st International Conference on Advanced Information Networking and Applications Workshops*, pp. 195-200 (2017)
5. Saito, N., Oda, T., Hirata, A., Hirota, Y., Hirota, M., Katayama, K.: Design and implementation of a DQN based AAV. In: Barolli, L., Takizawa, M., Enokido, T., Chen, H.-C., Matsuo, K. (eds.) *BWCCA 2020. LNNS*, vol. 159, pp. 321–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61108-8_32

6. Saito, N., Oda, T., Hirata, A., Toyoshima, K., Hirota, M., Barolli, L.: Simulation results of a DQN based AAV testbed in corner environment: a comparison study for normal DQN and TLS-DQN. In: Barolli, L., Yim, K., Chen, H.-C. (eds.) IMIS 2021. LNNS, vol. 279, pp. 156–167. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-79728-7_16
7. Saito, N., et al.: A Tabu list strategy based DQN for AAV mobility in indoor single-path environment: implementation and performance evaluation. *Internet Things* **14**, 100394 (2021)
8. Saito, N., Oda, T., Hirata, A., Yukawa, C., Kulla, E., Barolli, L.: A LiDAR based mobile area decision method for TLS-DQN: improving control for AAV mobility. In: Barolli, L. (ed.) 3PGCIC 2021. LNNS, vol. 343, pp. 30–42. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-89899-1_4
9. Wang, H., et al.: Automatic illumination planning for robot vision inspection system. *Neurocomputing* **275**, 19–28 (2018)
10. Zuxiang, W., et al.: Design of safety capacitors quality inspection robot based on machine vision. In: 2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS), pp. 1-4 (2017)
11. Li, J., et al.: Cognitive visual anomaly detection with constrained latent representations for industrial inspection robot. *Appl. Soft Comput.* **95**, 106539 (2020)
12. Ruiz-del-Solar, J., et al.: A Survey on Deep Learning Methods for Robot Vision. arXiv preprint [arXiv:1803.10862](https://arxiv.org/abs/1803.10862) (2018)
13. Matsui, T., et al.: FPGA implementation of a fuzzy inference based quadrotor attitude control system. In: Proceedings of IEEE GCCE-2021, pp. 691-692 (2021)
14. Saito, N., et al.: Approach of fuzzy theory and hill climbing based recommender for schedule of life. In: Proceedings of LifeTech-2020, pp. 368-369 (2020)
15. Ozera, K., et al.: A fuzzy approach for secure clustering in MANETs: effects of distance parameter on system performance. In: Proceedings of IEEE WAINA-2017, pp. 251-258 (2017)
16. Elmazi, D., et al.: Selection of secure actors in wireless sensor and actor networks using fuzzy logic. In: Proceedings of BWCCA-2015, pp. 125-131 (2015)
17. Elmazi, D., et al.: Selection of rendezvous point in content centric networks using fuzzy logic. In: Proceedings of NBiS-2015, pp. 345-350 (2015)
18. Zaeh, M.F., et al.: Improvement of the machining accuracy of milling robots. *Prod. Eng. Res. Devel.* **8**(6), 737–744 (2014)
19. Yukawa, C., et al.: Design of a fuzzy inference based robot vision for CNN training image acquisition. In: Proceedings of IEEE GCCE-2020, pp. 871-872 (2021)
20. Liang, Q., et al.: Interval type-2 fuzzy logic systems: theory and design. *IEEE Trans. Fuzzy Syst.* **8**(5), 535–550 (2000)
21. Mendel, J.M.: Interval type-2 fuzzy logic systems made simple. *IEEE Trans. Fuzzy Syst.* **14**(6), 808–821 (2006)
22. Yosinski, J., et al.: How transferable are features in deep neural networks? arXiv preprint [arXiv:1411.1792](https://arxiv.org/abs/1411.1792) (2014)
23. Zhuang, F., et al.: A comprehensive survey on transfer learning. *Proc. IEEE* **109**(1), 43–76 (2020)
24. Yukawa, C., et al.: Design of a robot vision system for microconvex recognition. In: Barolli, L., Kulla, E., Ikeda, M. (eds) International Conference on Emerging Internetworking, Data and Web Technologies, Lecture Notes on Data Engineering and Communications Technologies, vol. 118, pp 366-374. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95903-6_39
25. Yukawa, C., et al.: An intelligent robot vision system for recognizing micro-roughness on arbitrary surfaces: experimental result for different methods. In: Barolli, L. (eds) International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, vol. 496, pp. 221-229. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08819-3_22

26. Yukawa, C., et al.: Evaluation of a fuzzy-based robotic vision system for recognizing micro-roughness on arbitrary surfaces: a comparison study for vibration reduction of robot arm. In: Barolli, L., Miwa, H., Enokido, T. (eds) International Conference on Network-Based Information Systems, vol. 526, pp 230-237 (2022). https://doi.org/10.1007/978-3-031-14314-4_23



A Transportation Routing Method Based on A* Algorithm and Hill Climbing for Swarm Robots in WLAN Environment

Masahiro Niihara¹, Nobuki Saito², Chihiro Yukawa², Kyohei Toyoshima², Tetsuya Oda¹✉, Masaharu Hirota³, and Leonard Barolli⁴

¹ Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
t19j061nm@ous.jp, oda@ous.ac.jp

² Graduate School of Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
{t21jm01md,t22jm19st,t22jm24jd}@ous.jp

³ Department of Information Science, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
hirota@mis.ous.ac.jp

⁴ Department of Information and Communication Engineering, Fukuoka Institute of Technology, 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
barolli@fit.ac.jp

Abstract. The Vehicle Routing Problem (VRP) is an optimization problem that satisfies various constraints and minimizes the total route cost by multiple vehicles. In this paper, we propose a transportation routing method based on A* algorithm and *Hill Climbing (HC)* considering *WLAN connected Swarm Robots (WSRs)*. In addition, we compare the simulation results of the proposed method for the initial solution with/without the clustering *k*-means method considering the normal and uniform distributions of loads. The proposed method optimizes the transportation sequence and can decide the transportation route of each *WSR* considering WLAN connectivity.

1 Introduction

Automatic Guided Vehicles (AGVs)/Autonomous Mobile Robots (AMRs) [1, 2] can be used in warehouses and factories to reduce the employee burden, prevent human error and reduce labour costs. AGVs/AMRs can be more efficient by optimizing transportation routes considering pick-up and delivery loads. They also can be regarded as swarm robots [3–9] when operated on multiple units. If the swarm robots are connected to a Wireless Local Area Network (WLAN), the operators can manage and control the swarm robots constantly. In addition, the information necessary for collaborative tasks can be exchanged in order to detect the failure of robots. Satisfying various constraints and minimizing the total route cost by swarm robots for transportation correspond to the Vehicle Routing Problem (VRP) [10–12]. The decision of an optimal solution in VRP

is a *Non-deterministic Polynomial-time (NP)-hard* problem that requires exponential time to complete the computation of the optimal solution. There are some methods to solve the VRP including Local Search (LS) [13–16], Simulated Annealing (SA) [17–20] and Genetic Algorithm (GA) [21–24].

In this paper, we propose a transportation routing method based on A^* algorithm and *Hill Climbing (HC)* considering *WLAN connected Swarm Robot (WSR)*. In addition, we compare the simulation results by the proposed method that the initial solution with/without the clustering method based on k -means when the normal or uniform distribution of loads. The proposed method optimizes the transportation sequence and decides the transportation route of each *WSR* considering WLAN connectivity.

The structure of the paper is as follows. In Sect. 2, we show the proposed method. In Sect. 3, we show the simulation results. Finally, conclusions and future work are given in Sect. 4.

2 Proposed Method

In this section, we discuss the proposed method. Figure 1 shows the flowchart of the proposed method. The proposed method considers clustering based on k -means [25–27] to decide a more efficiently initial transportation route. We use the *HC* [28] for optimizing the transportation sequence and the A^* algorithm [29, 30] for deriving the transportation route of each *WSR* considering WLAN connectivity. The proposed method optimizes the transportation sequence and decides an efficient transportation route for each *WSR* that pick-up the loads in warehouses or factories considering WLAN connectivity. It is possible to centrally manage and remotely control the swarm robots constantly by guaranteeing the network connection with the swarm robots.

In the proposed method, all *WSRs* must be connected to one or more routers to consider network connectivity. In addition, the *WSRs* have their communication range and they can communicate when their communication range overlap. The routers are placed in a way that the maximum communication distance with the *WSRs* covers the target area.

2.1 Clustering Based K-Means for Transportation Routing

An efficient transportation can be achieved by collectively picking up of loads that are located near the placement of other loads. Therefore, the proposed method decides the initial transportation sequence for the transportation route of each *WSR* considering the k -means method which is one of the clustering methods.

The loads are clustered considering each placement of loads based on the k -means method into the number of *WSRs* used for transportation. It is expected that the optimal solution will be faster than using a random initial transportation sequence by using each cluster as the initial transportation sequence for each *WSR*.

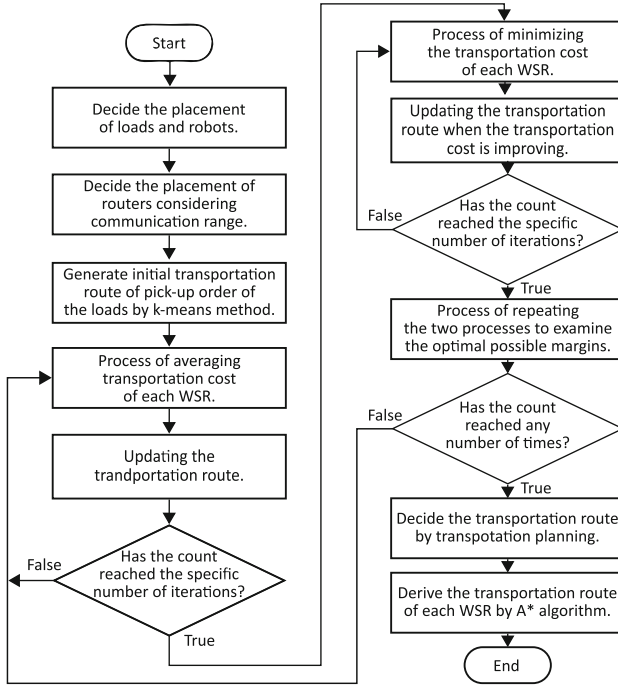


Fig. 1. Flowchart of proposed method.

2.2 Hill Climbing Based Optimization Method

The *HC* optimize the transportation sequence for each *WSR*. First, for each *WSR* is allocated a load that is based on each cluster by the clustering method as an initial transportation sequence and stores the placement of the loads in the transportation sequence of each *WSR*. Next, we perform the *HC* based optimization method [28]. The optimization method has two processes a process of averaging and a process of minimizing the transportation distance as the cost in the transportation sequence of each *WSR*. The transportation distance is the sum of the Euclidean distances between depot and loads, loads and loads.

In the actual environment, if the operating time of each *WSR* is biased when operating, that will be a negative influence, such as increasing the failures and wear/tear on *WSRs*. Therefore, the process of averaging prevents a bias in the transportation distance of each *WSR* and decreases the possibility that some *WSRs* will not be used. In addition, in the process of averaging, the mean squared error in the transportation distance of each *WSR* is minimized to reduce the transportation distance bias.

Table 1. Simulation parameters.

Parameters	Values
Area size $[(X_{min}, X_{max}), (Y_{min}, Y_{max})]$	$[(0.0, 200.0), (0.0, 200.0)]$
Number of routers [<i>unit</i>]	4
Number of <i>WSR</i> [<i>unit</i>]	4
Number of loads [<i>unit</i>]	40
Communication range of router [<i>unit</i>]	40.0
Communication range of <i>WSR</i> [<i>unit</i>]	30.0
Number of iteration [<i>times</i>]	1,000,000
Distribution of load placement [<i>unit</i>]	Normal/Uniform distribution

For each process, randomly choose one of the two that swap or pass over randomly selected loads from two randomly decided *WSRs*. The transportation sequence is update when the transportation distance of each *WSR* and the difference of transportation distance between *WSRs* are improved. If the above process repeats, the transportation sequence is not updated. After a specified number of times, the current best transportation sequence is decide as the optimal transportation sequence.

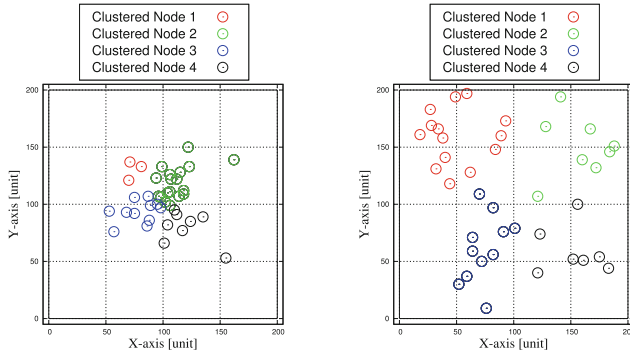
2.3 Transportation Routing Method Based on A^* Algorithm

After *HC* based optimization method decides the transportation sequence for pick-up of the loads by each *WSR*, the A^* algorithm decides the transportation route of each *WSR* considering the communication range and WLAN connectivity. The communication range is defined as the area when the Euclidean distance between the router and each *WSR* is shorter than the sum of the communication range of routers and the communication range of *WSR*.

3 Simulation Results

Table 1 shows the simulation parameters of the proposed method. The placement of router, *WSR* and depot are indicated by (*X*-axis, *Y*-axis). The initial placement of routers are (50.0, 50.0), (50.0, 150.0), (150.0, 50.0) and (150.0, 150.0), respectively. The initial placement of *WSRs* and depots are (90.0, 90.0), (90.0, 110.0), (110.0, 90.0) and (110.0, 110.0), respectively.

For the evaluation, we consider the total transportation distance of *WSR* and compare the perform once the proposed method without/with the clustering based *k*-means when the load placement has normal/uniform distribution. Figure 2 shows the visualization results by clustering method. Figure 3 shows the visualization result of the proposed method. Figure 3(a) and Fig. 3(b) show the optimized transportation route of each *WSR* when the initial solution is decided

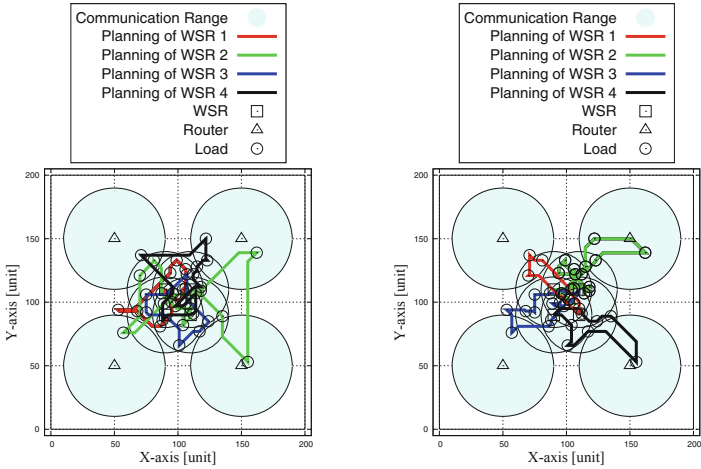


(a) The normal distribution of loads. (b) The uniform distribution of loads.

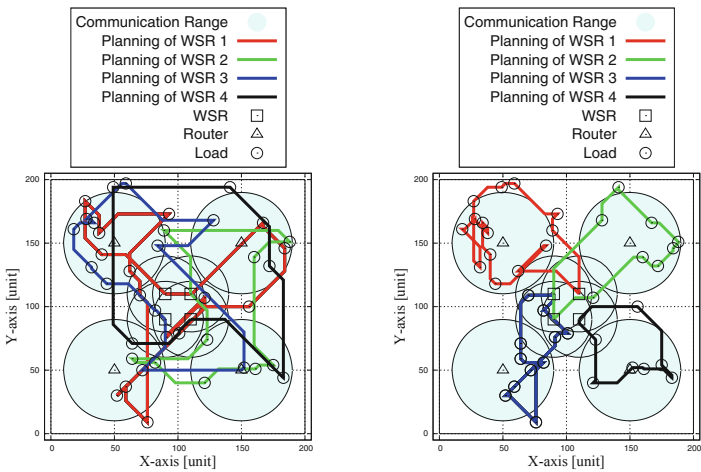
Fig. 2. Visualization results of clustering method.

without/with the clustering method and the load placement has normal distribution. Figure 3(c) and Fig. 3(d) show the optimized transportation route of each *WSR* when the initial solutions decided without/with the clustering method and the load placement has uniform distribution. The visualization results show that the proposed method can optimize the transportation sequence and decide an efficient transportation route for *WSRs* to pick up the loads and return to the depot.

From the simulation results, the total transportation distance without the clustering method is 1084.614 [unit] and the total transportation distance with the clustering method is 852.511 [unit] when we consider the normal distribution. Also, the total transportation distance without the clustering method is 2454.968 [unit] and the total transportation distance with the clustering method is 1331.780 [unit] when we consider the uniform distribution. The proposed method with clustering reduces the total transportation distance compared with the case without clustering by 21.399 [%] for the normal distribution and 45.751 [%] for the uniform distribution. Also, the visualization results show that the proposed method can reduce the transportation distance and decide an efficient route for pick-up the loads by each *WSR*.



(a) The normal distribution/without-Clustering-Method. (b) The normal distribution/with-Clustering-Method.



(c) The uniform distribution/without-Clustering-Method. (d) The uniform distribution/with-Clustering-Method.

Fig. 3. Visualization results of the proposed method.

4 Conclusions

In this paper, we proposed a transportation routing method based on the k -means method, HC and A^* algorithm considering WSR . The proposed method optimizes the transportation sequence and decides the transportation route of each WSR considering WLAN connectivity. The visualization results show that

the proposed method can reduce the transportation distance and decide an efficient route for pick-up the loads by each *WSR*. In the future, we would like to improve the proposed method.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number JP20K19793.

References

1. Sauer, T., et al.: Automatic track guidance of industrial trucks using self-learning controllers considering a priori plant knowledge. In: Proceedings of ICCAD-2021, pp. 1–8 (2021)
2. Jun, S., et al.: Pickup and delivery problem with recharging for material handling systems utilising autonomous mobile robots. *Eur. J. Oper. Res.* **289**, 1153–1168 (2021)
3. Albiero, D., et al.: Swarm robots in mechanized agricultural operations: a review about challenges for research. *Comput. Electron. Agric.* **193**, 106608 (2022)
4. Nguyen, L.A., et al.: Swarmathon: a swarm robotics experiment for future space exploration. In: Proceedings of ISMCR-2019, pp. B1-3-1–B1-3-4 (2019)
5. Oda, T., et al.: A deep Q-network based simulation system for actor node mobility control in WSANs considering three-dimensional environment: a comparison study for normal and uniform distributions. In: Proceedings of CISIS-2018, pp. 842–852 (2018)
6. Oda, T., et al.: Design and implementation of a simulation system based on deep q-network for mobile actor node control in wireless sensor and actor networks. In: Proceedings of IEEE AINA-2017, pp. 195–200 (2017)
7. Oda, T., et al.: Performance evaluation of a deep Q-network based simulation system for actor node mobility control in wireless sensor and actor networks considering different distributions of events. In: Proceedings of IMIS-2017, pp. 36–49 (2017)
8. Oda, T., et al.: Performance evaluation of a deep q-network based simulation system for actor node mobility control in wireless sensor and actor networks considering three-dimensional environment. In: Proceedings of INCoS-2017, pp. 41–52 (2017)
9. Toyoshima, K., et al.: A DQN based mobile actor node control in WSAN: simulation results of different distributions of events considering three-dimensional environment. In: Proceedings of EIDWT-2020, pp. 197–209 (2020)
10. Ancele, Y., et al.: Toward a more flexible VRP with pickup and delivery allowing consolidations. *Transp. Res. Part C: Emerg. Technol.* **128**, 103077 (2021)
11. Leng, K., Li, S.: Distribution path optimization for intelligent logistics vehicles of urban rail transportation using VRP optimization model. *IEEE Trans. Intell. Transp. Syst.* **23**(2), 1661–1669 (2022)
12. Chipuli, G.P., et al.: Analysis, design and reconstruction of a VRP model in a collapsed distribution network using simulation and optimization. *Case Stud. Transp. Policy* **9**, 1440–1458 (2021)
13. Maximo, V.R., Nascimento, M.C.V.: A hybrid adaptive iterated local search with diversification control to the capacitated vehicle routing problem. *Eur. J. Oper. Res.* **294**, 1108–1119 (2021)

14. Voigt, S., et al.: Hybrid adaptive large neighborhood search for vehicle routing problems with depot location decisions. *Comput. Oper. Res.* **146**, 105856 (2022)
15. Hirata, A., et al.: Improvement of NMR-reduction method by local search for optimization of number of mesh routers in WMNs. In: *Proceedings of NBiS-2022*, pp. 66–77 (2022)
16. Hirata, A., et al.: Simulation results of CCM based HC for mesh router placement optimization considering two islands model of mesh clients distributions. In: *Proceedings of EIDWT-2021*, pp. 180–188 (2021)
17. Motaghedi-Larijani, A.: Solving the number of cross-dock open doors optimization problem by combination of NSGA-II and multi-objective simulated annealing. *Appl. Soft Comput.* **128**, 109448 (2022)
18. Rao, T.S.: A simulated annealing approach to solve a multi traveling salesman problem in a FMCG company. *Mater. Today: Proc.* **46**, 4971–4974 (2021)
19. Hirata, A., et al.: A simulation system for mesh router placement in WMNs considering coverage construction method and simulated annealing. In: *Proceedings of BWCCA-2021*, pp. 78–87 (2021)
20. Sakamoto, S., et al.: Implementation of an intelligent hybrid simulation system for node placement problem in WMNs considering particle swarm optimization and simulated annealing. In: *Proceedings of IEEE AINA-2017*, pp. 697–703 (2017)
21. da Costa, P.R.O., et al.: A genetic algorithm for a green vehicle routing problem. *Electron. Notes Disc. Math.* **64**, 65–74 (2018)
22. Xin, L., et al.: Logistics distribution route optimization based on genetic algorithm. *Nat.-Insp. Comput. Web Intell.* **2022**, 8468438 (2022)
23. Oda, T., et al.: Evaluation of WMN-GA for different mutation operators. *Int. J. Space Based Situat. Comput.* **2**(3), 149–157 (2012)
24. Oda, T., et al.: WMN-GA: a simulation system for WMNs and its evaluation considering selection operators. *J. Ambient Intell. Human. Comput.* **4**(3), 323–330 (2013)
25. Fatemi-Anaraki, S., et al.: A hybrid of K-means and genetic algorithm to solve a bi-objective green delivery and pick-up problem. *J. Ind. Prod. Eng.* **39**, 146–157 (2020)
26. Silva, C.E., et al.: Route scheduling system for multiple self-driving cars using K-means and bio-inspired algorithms. *Eng. Appl. Neural Netw.* **1600**, 27–39 (2022)
27. Obukata, R., et al.: Design and evaluation of an ambient intelligence testbed for improving quality of life. *Int. J. Space Based Situat. Comput.* **7**(1), 8–15 (2017)
28. Liu, H.D., et al.: A novel photovoltaic system control strategies for improving hill climbing algorithm efficiencies in consideration of radian and load effect. *Energy Conv. Manag.* **165**, 815–826 (2018)
29. Bagheri, S.M., et al.: An A-star algorithm for semi-optimization of crane location and configuration in modular construction. *Autom. Constr.* **121**, 103447 (2021)
30. Bai, R., et al.: Analytics and machine learning in vehicle routing research. *Int. J. Prod. Res.*, 1–27 (2021)



Simulation of Choice of Residence for Working Women

Risa Takata^(✉), Shiori Koga, and Kaoru Fujioka

Fukuoka Women's University, 1-1-1 Kasumigaoka,
Higashi-ku, Fukuoka 813-8529, Japan
19ue029@mb2.fwu.ac.jp, kaoru@fwu.ac.jp

Abstract. In recent years, Japan's declining birthrate and aging population have increased the need for new urban planning through sustainable compact city policies, and have also increased the importance of women in the workforce. Based on the above background, this study focuses on the choice of residence among working women. We simulate land use patterns in a virtual city, assuming that women choose a place to live according to the factors that are important to them, such as proximity to a train station and a nursery school or school-age childcare. These factors are included based on the results of a questionnaire survey on the items that women consider important for choosing a place to live. We analyzed the location of the houses in the city when the simulation converged. Our results showed that the location of the houses varies depending on the household.

1 Introduction

In recent years, there has been an increasing need for urban planning suitable for new lifestyles in Japan, such as compact city policies for sustainable development in response to the declining the birthrate and aging population, and urban development that takes into account changes brought about by the COVID-19 pandemic. On the other hand, changes in the social environment surrounding women and changes in women's attitudes toward work have made it even more important to realize a society in which women can easily balance work with housework and childcare.

In our previous study [1], several models were created with different initial arrangements of facilities, including a virtual city model with randomly arranged facilities and a virtual city model with workplaces concentrated in the center. The results showed that the location of the workplace affects the choice of residence for both single and family households. The simulation results of our previous study [1] and the numerical analysis from another study by Waddel [2] indicate that distance from the workplace and commuting time are important factors influencing residential location decisions. According to another study by Yui [3], women tend to place more importance on security and childcare services than men when choosing a place to live, and it has been noted that there are gender

differences in the items that are important when choosing a place to live. Yet another previous study found that women’s commuting burdens affect where couples live [5]. Yui [3] conducted an online questionnaire survey on the items that working women consider important when choosing a place to live and the results were divided into three groups: single women (referred to as *singles*), women in dual-earner households without children (referred to as *dinks*), and women in dual-earner households with children (referred to as *family*).

In the present study, for the three household types, singles, dinks, and family, we create a simulation model of residence choice based on the results of the questionnaire in [4], and the characteristics of the residences of each household are analyzed.

2 Methods

We consider a city of small to medium scale, with a space of 50×50 cells, each cell is assumed to have a side length of 50 to 100 m, and only one agent can be placed in each cell. In this study, eight types of agents are created in the model space: *house* (resident), *office* (workplace), *market* (store), *park* (park), *nursery* (nursery), *station* (station), *school* (school), and *empty* (empty space) as shown in Table 1. Six of these types of agents, excluding house and empty, are denoted as *institution* agents.

House agents are classified into three types of households, denoted by *singles*, *dinks*, and *family*, according to *housing expenses*, which indicates the maximum amount of spending available for housing. A housing expense is set by a uniform random number between 0 and 90. Referring to the survey results by Japan’s Ministry of Health, Labour and Welfare [6] regarding the housing cost burden ratio of households in their 30s, for a house agent with housing expense he , if $0 \leq he < 30$ (resp. $30 \leq he < 60$, $60 \leq he < 90$), the agent belongs to family (resp. dinks, singles).

Table 1. Types of agents and the number of agents.

Types of agents	Number of agents
House (resident)	1000
Office (workplace)	54
Market (store)	30
Nursery (nursery)	15
School (school)	15
Park (park)	10
Station (train station)	1
Empty (empty space)	1375

In order to analyze the effects of the placement of institutions on residential placements, we set two models: one in which the institution agents are concentrated in the center of the space (called *central placement model*), as shown in Fig. 1a, and the other in which they are scattered throughout the space (called *random placement model*), as shown in Fig. 1b. The institution agents remain stationary throughout the entire simulation.

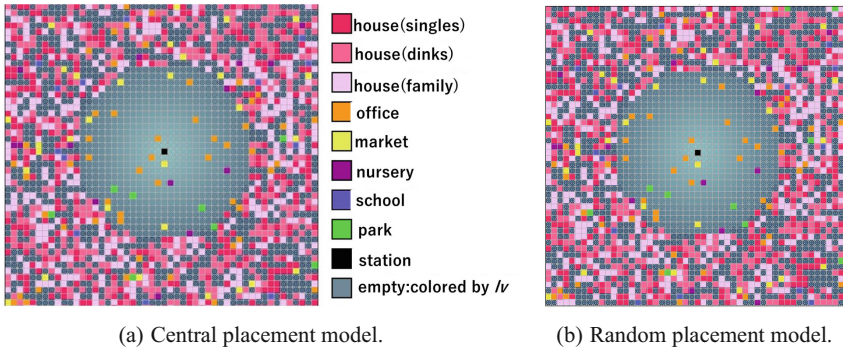


Fig. 1. The field of the model at the initial step

We set the land value of each cell in the space according to its distance from the station agent placed at position (25, 25). Approximately half of the cells in the space (1245 cells), are within 14 units of distance from the station agent, and these cells are referred to as being in the *downtown area*. A cell in the central area is set so that the land value decreases with the distance from the station agent. Specifically, *land value* lv is set as $lv = 90 - p \cdot d$ for the distance d to the station agent. Based on the results from a preliminary experiment, the parameter p is set as 5. For the other cells, referred to as being in the *uptown area*, the land value is kept constant regardless of the distance from the station agent.

The items that each household considers important when selecting a place to live are set with reference to the results of a questionnaire from a previous study [4]. Questionnaire results unrelated to our model (e.g., housing facilities) are excluded in this study. Based on the items that each household considered important when choosing a place to live and the percentage of respondents who answered that the item was important to them in the survey results, we determine the probability that a house agent will take these items into account when moving in our model. For example, our model sets the probability that a single house agent values being near a train station at 71%, since 71% of single women responded in the questionnaire that they value being close to a train station.

Table 2. Residential choice of working women [4]

Singles	%	Family	%
Station	71	Proximity to train station	53
Land price/rent	55	Land price/rent	49
Proximity to work	54	Proximity to husband's work	29
Convenience of stores and facilities	16	Proximity to parents	24
Quiet environment	11	Convenience of stores and facilities	22
Proximity to acquaintances and friends	11	Proximity to work	17
		Familiar environment	16
Dinks	%		
Land price/rent	46		
Proximity to train station	32		
Proximity to parents	27		
Familiar environment	24		
Proximity to husband's work	18		
Children's educational environment	18		
Convenience of stores and facilities	16		
Nursery school or school-age childcare	12		
Proximity to work	10		
Environment in proximity to nature	10		

The following shows how the questionnaire items in Table 2 are reflected in the model.

- Proximity to train station: Distance to the station agent.
- Land price/rent: Land value lv .
- Proximity to work: Distance to a certain office agent that is randomly selected at the beginning of the simulation.
- Convenience of stores and facilities: Distance to the nearest market agent.
- Quiet environment: Distance to the station agent multiplied by a negative value.
- Proximity to acquaintances and friends: Distance to a certain single house agent that is randomly selected at the beginning of the simulation. Even if the selected agent moves, consider the distance to that agent.
- Proximity to parents: Distance to a certain family agent that is randomly selected at the beginning of the simulation. Even if the selected agent moves, consider the distance to that agent.
- Familiar environment: Distance to a certain family agent that is randomly selected at the beginning of the simulation. Even if the selected agent moves, consider the distance to that agent.
- Children's Educational Environment: Distance to the nearest school agent.
- Securing daycare and childcare for school children: Distance to the nearest nursery agent.

- Environment in proximity to nature: Distance to the nearest park agent.

At each step, a house agent calculates the dissatisfaction level by adding up the housing expense he and the above-mentioned factors. Each house agent checks whether a randomly selected empty agent satisfies the following two conditions, and moves by exchanging its location with the empty agent if it satisfies the conditions:

1. The selected empty agent satisfies $lv < he$, which means that the land value is less than the housing expense of the house agent that is the current focus.
2. The dissatisfaction level of the selected empty agent denoted by D_e is lower than the current dissatisfaction level denoted by D , i.e., $D > D_e$, which means that the dissatisfaction as a residential area is compared between the present location and the selected location.

Each house moves according to the above rules, and the simulation is terminated when all house agents satisfy $D \leq D_e$. The simulation is executed 100 times each for the central placement model and the random placement model.

3 Experiments and Analysis

The field of the model at convergence for the central placement model and the random placement model is shown in Fig. 2a and Fig. 2b, respectively. In the central placement model, most of the single house agents are located in the downtown area, while most of the dinks and family house agents are located on the boundary between the downtown and the uptown area. In the random placement model, house agents are more sparsely distributed than in the central placement model and are distributed throughout the entire space. The number of house agents in the downtown area in the random placement model is smaller than that in the central placement model.

From the results of 100 simulations, the mean, median, and variance of the distance from the station at the end of the simulation are calculated. The results for the central placement model and the random placement model are summarized in Table 3. The random placement model has larger values than the central placement model for all means, medians, and variances. The values in Table 3 as well as the fields represented in Fig. 2a and Fig. 2b show that the central placement model is more successful in compacting and centralizing the city.

We obtain the distance from the station to all house agents at the end of the simulation as a rounded integer value, then calculate the corresponding number of house agents for each distance to the station. The average number of house agents is obtained for each distance to the station agent from 100 iterations of the simulation. Figure 3a (resp. Fig. 3b) shows the average number of house agents on the vertical axis and the distance from the station agent on the horizontal axis for the central replacement model (resp. random replacement model).

Both models show a significant number of house agents near the boundary between downtown and uptown. In the central placement model, the number of

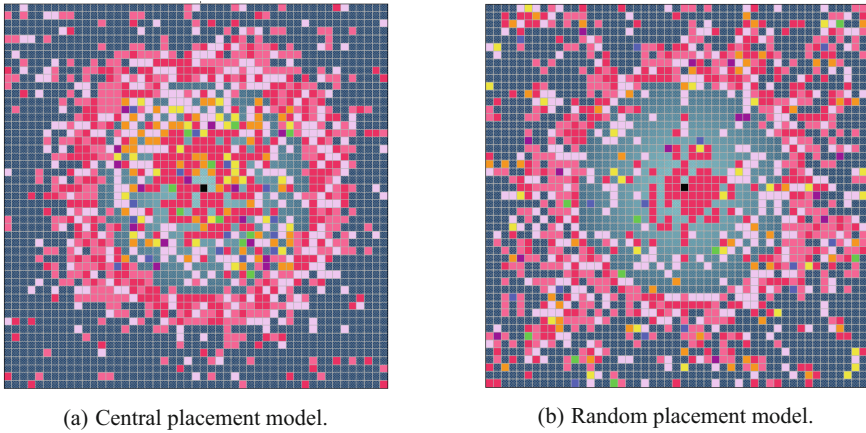


Fig. 2. The field of the model at convergence.

Table 3. The distance from the station to all house agents at the end of the simulation consisting of 100 trials.

	Mean	Median	Variance
Central placement model	15.8	16.2	33.6
Random placement model	17.6	17.5	40.4

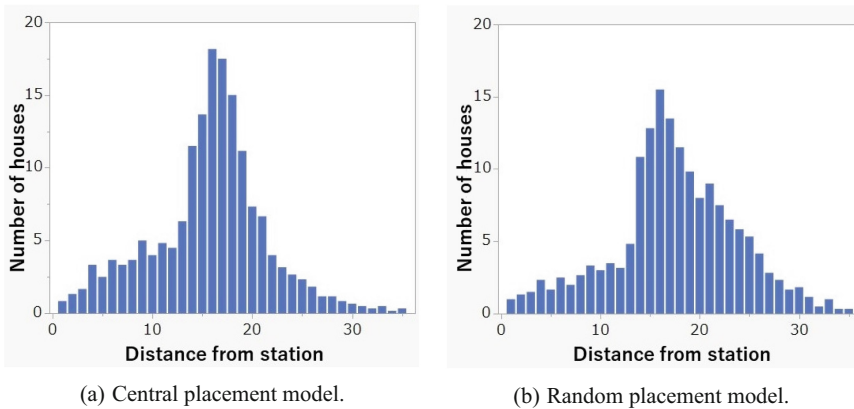


Fig. 3. Distance from the station agent and the number of house agents per household

house agents in the downtown area per unit distance to the station tends to be higher than in the random placement model. In the uptown area, the number of house agents tends to decrease rapidly with distance from the station in the central placement model, while the decrease is slower in the random placement model.

Based on data shown in Table 3, Table 4 summarizes the result for each household. The mean and median indicate that the distance from the station agent is smaller in the order of singles, dinks, and family for the two models. On the other hand, variance is larger in the order of singles, dinks, and family.

Table 4. The distance from the station to all house agents at the end of the simulation for 100 trials per household.

		Mean	Median	Variance
Central placement model	Singles	13.5	15.0	47.4
	Dinks	15.4	15.3	26.5
	Family	18.5	18.0	14.3
Random placement model	Singles	15.7	15.8	58.4
	Dinks	17.3	16.8	34.2
	Family	19.8	19.1	19.9

Based on data shown in Fig. 3a and Fig. 3b on the distance to the station agent and the number of house agents, Fig. 4a and Fig. 4b show the results for each household, respectively. There are two distribution peaks for singles and dinks house agents, but there is only one peak for family house agents. These peaks are more prominent in the central placement model than in the random placement model. Single house agents can be seen at distances both short and far from the station, while family house agents are rarely found at a distance of less than 11 units from the station.

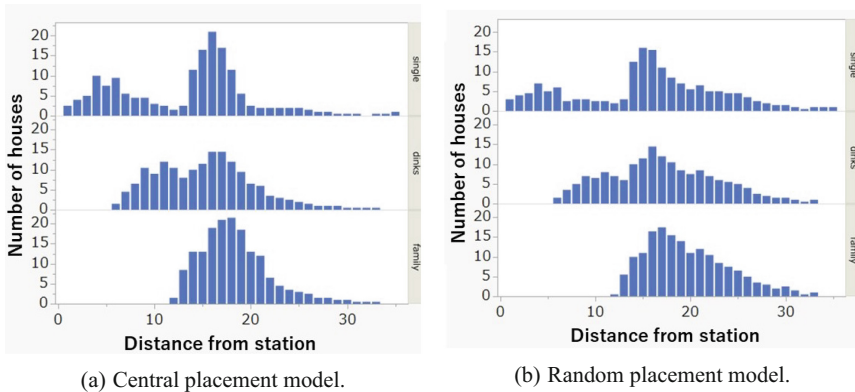


Fig. 4. Distance from the station agent and the number of house agents

4 Conclusion

In this study, a city simulation model for land use patterns was developed based on the results of a questionnaire concerning items that women consider when selecting a place to live. The simulation results indicate that the central placement model, in which various facilities are located in the center of the city, promotes urban compactness. In addition, different households show different characteristics with respect to the choice of place of residence.

Our future work is to analyze which factors have a significant impact on the choice as to where to live, such as land cost and proximity to train stations. In addition, it will also be important to investigate the optimal layout of facilities in a city.

References

1. Fujioka, K., Nozato, A.: Development of urban simulation model for aiding compact city policies. *Environ. Inf. Sci.* **34**, 79–84 (2019). (in Japanese)
2. Waddell, P., Bhat, C., Eluru, N., Wang, L., Pendyala, R.M.: Modeling interdependence in household residence and workplace choices. *J. Transp. Res. Board* **2003**(1), 84–92 (2007)
3. Yui, Y.: Housing issues from a gender perspective. *Q. J. Household Econ.* **69**, 37–46 (2006). (in Japanese)
4. Yui, Y., Wakabayashi, Y., Nakazawa, T., Kamiya, H.: Residential choices of working women in urban space. *E-J. GEO* **2**(3), 139–152 (2007). (in Japanese)
5. Ora, F., Clifford, K.R.: A model of workplace and residence choice in two-worker households. *Reg. Sci. Urban Econ.* **27**(3), 241–260 (1997)
6. Ministry of Health, Labour and Welfare. Analysis of national income and living conditions and others. https://www.mhlw.go.jp/file/05-Shingikai-12601000-Seisakutoukatsukan-Sanjikanshitsu_Shakaihoshoutantou/0000184114.pdf



Constructing and Reconstructing Characters Using Gaussian Process Regression

Jinya Yano¹ and Hiroyuki Fujioka²(✉)

¹ Graduate School of Engineering, Fukuoka Institute of Technology,
3-30-1 Wajiro-Higashi, Higashi-ku, Fukuoka 811-0295, Japan
mjm22103@bene.fit.ac.jp

² Department of System Management, Fukuoka Institute of Technology,
3-30-1 Wajiro-Higashi, Higashi-ku, Fukuoka 811-0295, Japan
fujioka@fit.ac.jp

Abstract. In this paper, we consider the problem of constructing characters from a set of measured data using some electronic devices - such as pen tablet and electronic white board, etc. For such a problem, the Gaussian Process Regression is employed, where the Gaussian kernel is used as kernel function. Then, the character is then obtained as the average of posterior distribution based on the Gaussian process model. In particular, it is shown that transforming the typeface into a running style is available by adjusting some hyper-parameter of the kernel function. The design examples are included.

1 Introduction

Writing is one of the essential tasks in even the today's digitalized world [1]. In the writing task, human produces characters by moving a writing device (e.g. pen) on a writing plane (e.g. paper and display) continuously in both time and space. Based on this observation, Takayama and Kano et al. [2] have developed a novel font design method called as 'dynamic font'. Unlike the ordinary design methods -such as dot matrix, outline vector, and skeleton vector methods, etc. [3], the idea was to introduce a virtual writing device and virtual writing plane, and characters are formed as the result of three-dimensional (3-D) motion of device. Moreover, by designing the motion as a weighted sum of time-shifted B-splines [4], the motion is represented by the sequence of 3-dimensional weight vectors. In other words, a character, a motion of the writing device, and weight vectors of B-splines are all regarded as the same in representation. This fact constituted a formal theory for handling characters. In particular, we have developed a scheme for generating cursive characters by employing control-theoretic B-spline function approximation [5]. Very recently, one of authors in [6] has also developed a scheme for modeling hairy-brush characters as seen in Japanese calligraphy, by introducing a deep learning approach.

However, in using the dynamic font, when generating characters with complex shapes such as handwritten characters, it is necessary to maintain a large size of B-spline weight vectors. For such a problem, one possible improvement would be to use

sparse modeling [7]. But, this is not realistic because it requires the preparation of dictionaries for a wide variety of characters in advance.

In this study, we develop a new method for constructing and reconstructing characters. But, the big differences with work in [5] are as follows:

- (i) Use of the Gaussian process regression (GPR) [8] for constructing characters.
- (ii) Reconstructing the style of characters and generating ink-effects on characters are available using GPR.

In (i), supposing that handwriting motion in some writing are measured using a pen tablet, then we construct characters using GPR from such a measured data. Here, the Gaussian kernel is used as the kernel function. Then, it is shown that characters are obtained as the mean of the posterior distribution based on the Gaussian process model. As for (ii), the fundamental operations –such as resizing and translating characters, are introduced. Then, it is shown that an operation to transform the typeface to a writing style can be defined by adjusting the hyper-parameter of the kernel function. Furthermore, by utilizing the variance of the posterior distribution, it is possible to express “blurring” as seen in brush strokes. The usefulness of this method is shown by a design example.

The paper is organized as follows. In Sect. 2, we present the method for constructing characters using Gaussian process regression. Then, the method for manipulating characters is described in Sect. 3. In Sect. 4, a design example is included, and concluding remarks are given in Sect. 5.

2 Constructing Characters Using Gaussian Process Regression

We first present the method for constructing characters using Gaussian process regression [8].

Now, we suppose that handwriting motions are measured by using pen-tablet and they are stored as a set of data \mathcal{D} ,

$$\mathcal{D} = \{(x_i, y_i) : x_i \in [x_{\text{ini}}, x_{\text{fin}}], y_i \in \mathbf{R}^2, i = 1, 2, \dots, N\}, \quad (1)$$

where the mean of y is normalized to be 0.

We then consider to represent a set of data \mathcal{D} using an idea of Gaussian Process Regression (GPR). Namely, we construct a model to estimate a function $f(x)$ with an input $x \in \mathbf{R}^+$ (i.e. time) and writing motion $y \in \mathbf{R}^2$ defined as

$$y = [X, Y]^T \in \mathbf{R}^2. \quad (2)$$

For simplicity, we here assume that each element in $y(x) \in \mathbf{R}^2$, i.e. $X(x)$ and $Y(x)$, is constructed independently in the sequel, and using notational abuse, $y(x)$ is treated as scalars with the understanding that they represent one of the two elements.

Here, we suppose that the function f is generated from a Gaussian process, i.e.

$$f \sim \text{GP}(0, k(x, x)).$$

Letting \mathbf{y} be a vector defined as

$$\mathbf{y} = [y_1, y_2, \dots, y_N]^T, \quad (3)$$

we have $y \sim N(\mathbf{0}, \mathbf{K})$ using a kernel matrix $\mathbf{K} \in \mathbf{R}^{N \times N}$ given by

$$\mathbf{K} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) \end{bmatrix}. \quad (4)$$

Here, the kernel function $k(x_i, x_j)$ is a Gaussian kernel, i.e.

$$k(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{\theta}\right), \quad (5)$$

where θ denotes a hyper-parameter.

Letting $p(y^* | x^*, \mathcal{D})$ be a posterior distribution of y^* corresponding to some time $x^* \in [x_{\text{ini}}, x_{\text{fin}}]$, we obtain

$$p(y^* | x^*, \mathcal{D}) = N(\mathbf{k}_* \mathbf{K}^{-1} \mathbf{y}, k_{**} - \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_*), \quad (6)$$

where \mathbf{k}_* and k_{**} are given by

$$\mathbf{k}_* = [k(x^*, x_1) \ k(x^*, x_2) \ \cdots \ k(x^*, x_N)]^T, \quad (7)$$

$$k_{**} = k(x^*, x^*). \quad (8)$$

Then, characters are constructed as an expectation of posterior in (6), i.e.

$$E(y^* | x^*, \mathcal{D}) = \mathbf{k}_* \mathbf{K}^{-1} \mathbf{y}, \quad (9)$$

where x^* is at a point sampled appropriately in $[x_{\text{ini}}, x_{\text{fin}}]$.

Remark 1. Note here that the size of kernel matrix \mathbf{K} in (4) increases as the number of data, i.e. N , becomes large. Then, the computational complexity of \mathbf{K}^{-1} in (9) increases. For such an issue, some approximation algorithms can be used [9].

3 Reconstructing Characters

We are now in the position to develop a scheme for reconstructing characters in Sect. 2.

Once we compute K^{-1} in (9), the fundamental operations - such as resizing and translation of characters, can be achieved by linear transformation on \mathbf{y} in (9) as follows.

We here introduce a fundamental operator $\mathcal{T}(\cdot; \cdot; \cdot)$ for resizing and translating characters such that, for a given \mathbf{y} and scalar $a, c \in \mathbf{R}$,

$$\mathcal{T}(\mathbf{y}; a; c) = [a \cdot y_1 + c, a \cdot y_2 + c, \dots, a \cdot y_N + c]^T. \quad (10)$$

Then, the fundamental operations - such as resizing and translation of characters, can be achieved by linear transformation on \mathbf{y} in (9). When we want to resize and translate characters constructed by (9), we have only to replace \mathbf{y} with $\mathcal{T}(\mathbf{y}; a; c)$ and compute $E(y^*|x^*, \mathcal{D})$, i.e.

$$E(y^*|x^*, \mathcal{D}) = \mathbf{k}_* \mathbf{K}^{-1} \mathcal{T}(\mathbf{y}; a; c). \quad (11)$$

The operation for rotating characters can be also defined in the same manner.

On the other hand, θ in (5) is a parameter corresponding to how much of the estimated value reflects nearby points. As θ becomes large, the estimated y^* becomes smoother. Thus, the style of characters can be transformed to some cursive ones. In addition, the blurring that can be seen in brush strokes can be depicted by superimposing the characters generated by (9).

Remark 2. Note here that the inverse matrix \mathbf{K}^{-1} in (9) must be recalculated when we set some different value of θ , where some approximation algorithms can be used as in Remark 1.

4 A Design Example

We show an example of characters constructed and reconstructed by using the scheme presented in Sects. 2 and 3.

Figure 1 illustrates a set of measured data \mathcal{D} for the handwritten character “hello” on XY plane. Here, the dataset was measured and stored by using the system in Fig. 2. This measurement system consists of PC with Linux OS and pen-tablet device. By this system, we measure and store handwriting input on the display as a dataset \mathcal{D} consisting of 2-dimensional position on XY-plane at the sampling rate 0.6 [s].

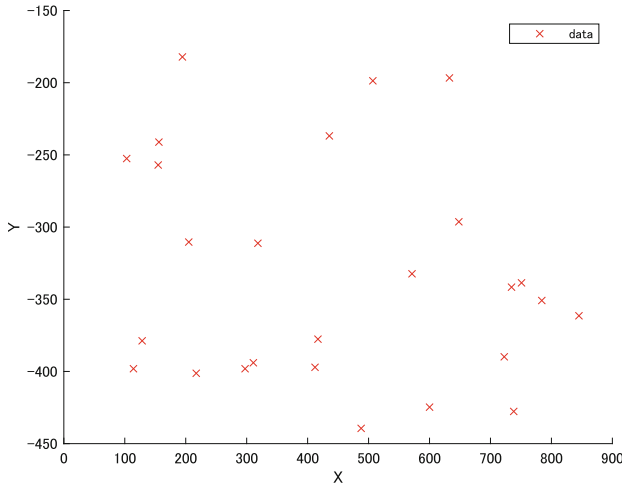


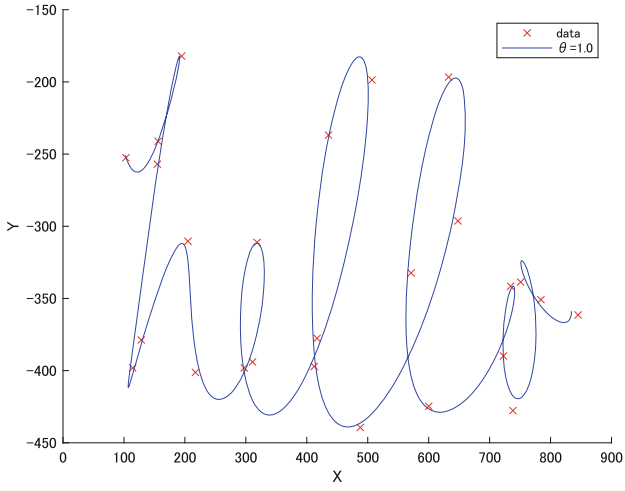
Fig. 1. A measured dataset \mathcal{D} for the handwritten character “hello”.



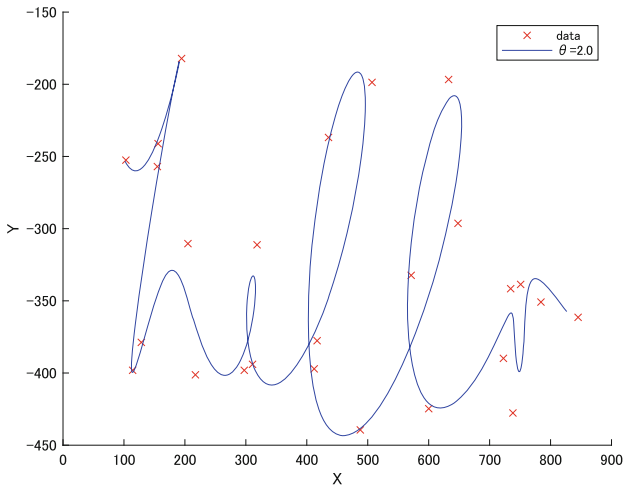
Fig. 2. Overview of system for measuring handwriting motion.

In Fig. 3, we show a design example of characters constructed from a dataset in Fig. 1. We here set x_{ini} and x_{fin} as $x_{ini} = 0$ and $x_{fin} = 15$, respectively. Hence, N is set as $N = 26$. Figures 3(a) and (b) show the results for the cases where θ in (5) is set as $\theta = 1.0$ and $\theta = 2.0$, respectively. The cross marks indicate the two-dimensional position $y = [X, Y]^T$ of the measured handwriting motion. We here observed that a large value of θ can deform the style well like cursive script.

On the other hand, Fig. 4 shows the results of superimposing writing motions 10 times (Fig. 4(a)) and 1000 times (Fig. 4(b)) by $p(y^*|x^*, \mathcal{D})$, where θ is set as $\theta = 1.0$. Then, we observed that the style of reconstructed characters can be expressed like a brush stroke.

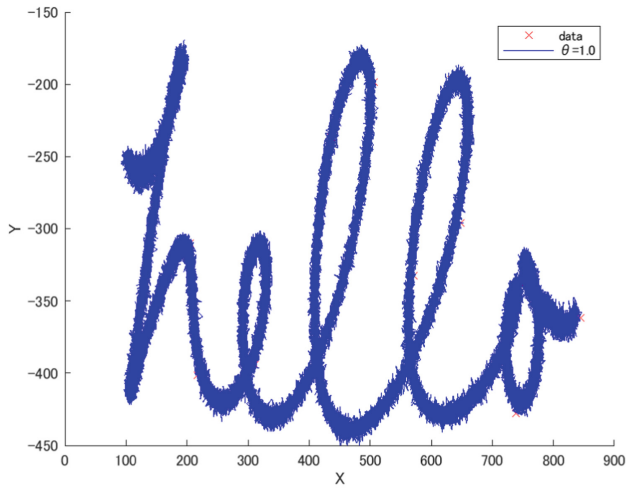


(a) $\theta = 1.0$

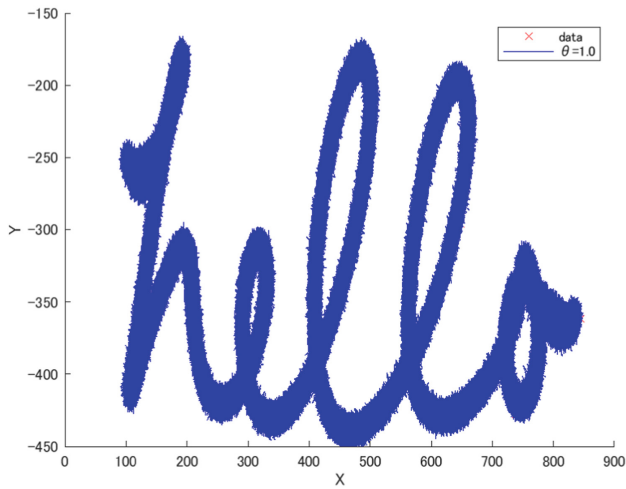


(b) $\theta = 2.0$

Fig. 3. An example of constructed characters “hello”.



(a) 100 times



(b) 1000 times

Fig. 4. The results of superimposing writing motions (a) 10 times and (b) 1000 times by $p(y^*|x^*, \mathcal{D})$.

5 Concluding Remarks

In this study, we developed a method for constructing and reconstructing characters using Gaussian process regression. First, we presented a method for constructing characters from a measured data on human handwriting. For constructing characters, we employed Gaussian process regression, where the Gaussian kernel is used as the kernel

function. Then, we showed that characters are obtained as the mean of the posterior distribution based on the Gaussian process model. In such a scheme, we define the basic operations to handle font sizes and locations. In addition, it is shown that an operation to transform the typeface to a writing style is available by adjusting the hyper-parameter of the kernel function. Also, by utilizing the variance of the posterior distribution, it was possible to express “blurring” as seen in brush strokes. The usefulness of this method was demonstrated by a design example.

References

1. Kurtenbach, G.: Pen-based computing, XRDS: crossroads. *ACM Mag. Students Future Interact.* **16**(4), 14–20 (2010)
2. Takayama, K., Kano, H., et al.: Dynamic font: a new representation technology. *Fujitsu Sci. Tech. J.* **32**(2), 192–202 (1996)
3. Uehara, T.: Current technology and problems in computer font. *Trans. Inf. Process. Soc. Japan* **31**(11), 1570–1580 (1990). (in Japanese)
4. de Boor, C.: *A Practical Guide to Splines*. Springer, New York (1978)
5. Fujioka, H., Kano, H., Nakata, H., Shinoda, H.: Constructing and reconstructing characters, words, and sentences by synthesizing writing motions. *IEEE Trans. Syst. Man Cybern. Part A* **36**(4), 661–670 (2006)
6. Xie, Z., Fujioka, H., et al.: Constructing and reconstructing dynamic font-based hairy brush characters using control-theoretic approach. In: *Proceedings the 21th IFAC World Congress, Berlin, Germany, 12–17 July 2020*, 6 p (2020)
7. Mieno, Y., Fujioka, H., Kano, H.: Data compression of digital-ink with pen-slips using multi-level L1 smoothing splines. In: *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics, Hong Kong, 9–12 October 2015*, pp. 1787–1792 (2015)
8. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer, New York (2006)
9. Akaho, S.: Introduction to Gaussian process regression. *Syst. Control Lett.* **62**(10), 390–395 (2018). (in Japanese)



Proposal of Disaster Prevention Training System Using Mixed Reality Space

Takahiro Uchiya^(✉) and Kazuki Akita

Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555, Japan
t-uchiya@nitech.ac.jp, k.akita.179@nitech.jp

Abstract. The probability of a large Nankai Trough earthquake occurring within 30 years, perhaps causing extensive damage, has been estimated recently as 70–80%. Although disaster prevention drills are held by large organizations once annually, younger people tend to have low awareness of disaster prevention. In response to a recent trend of people neglecting disaster education, research has been conducted to raise disaster prevention awareness using virtual reality (VR) to simulate disaster sites and experiences. This study proposes and examines disaster prevention training system materials that incorporate mixed reality (MR), which reproduces VR on a real-space basis. This system materials are aimed at improved confidence after training and reduced risk of motion sickness (VR sickness), which earlier studies of VR training materials have identified as difficulties. As described herein, after the contents of the proposed method are explained, the effectiveness of the proposed method is demonstrated using evaluation experiments.

1 Introduction

To a greater degree than most countries, disasters such as earthquakes occur frequently in Japan. Particularly, the Great East Japan Earthquake of March 2011 wrought considerable damage. In recent years, the probability of a magnitude 8–9 Nankai Trough Earthquake occurring within 30 years has been estimated as 70–80% [1]. Of course, 30 years is not a short period of time for humans. Moreover, it is difficult to predict exactly when an earthquake will occur. Also, it is impossible to live in constant vigilance against natural disasters such as earthquakes and fires. For those reasons, disaster drills are held once annually by large organizations. However, some disaster drill participants remain unaware of the necessity of disaster drills or have low awareness of disaster prevention. Such tends to be the case particularly for people of younger generations. In response to this trend away from disaster prevention education, studies have been undertaken to raise disaster prevention awareness using virtual reality (VR) [2] to experience simulated events at disaster sites. This paper proposes disaster prevention training system materials incorporating mixed reality (MR) [3], based on the idea that MR, which reproduces VR on a real-space basis, can provide “reduced risk of motion sickness” in addition to “improved confidence after training”, which were salient difficulties with VR materials examined in earlier studies, while retaining positive aspects of the VR experience.

2 Earlier Research

2.1 Overview

Tanimoto et al. [4] proposed a disaster prevention training system using VR to address the recent trend toward neglect of disaster prevention education. The use of VR can provide immersive and realistic experiences of disasters that are not provided by conventional teaching materials. Moreover, VR can present disasters that cannot be seen or experienced without being actually encountered. Although simulation of a disaster environment by burning things to simulate a fire is not impossible, it is unrealistic to do so repeatedly. In this respect, VR presents the benefit of requiring no use of consumable items. Moreover, VR can ensure the safety of the training participants. The proposed system comprises three parts: a VR evacuation drill, a VR fire drill, and a VR comprehensive drill. The attention, relevance, confidence, and satisfaction (ARCS) model from educational engineering is incorporated in drills to reduce disengagement from disaster education and reduced disaster awareness, which are regarded as difficulties inherent in disaster education.

2.2 ARCS Model

As proposed by John Keller in 1983, the ARCS model helps learners to improve their motivation to learn. The model is a framework that organizes learning motivation problems and measures into the four categories of attention, relevance, confidence, and satisfaction. The framework also provides motivational policies and procedures for designing motivation for each category.

Indicators for each item are presented below.

- Attention: Did the learner devote attention to the contents of the material?
- Relevance: Did learners find the contents of the material rewarding?
- Confidence: Are learners confident that they can actually use the material?
- Satisfaction: Does the learner feel a sense of accomplishment after learning the material?

The Instructional Materials Motivation Survey (IMMS) scale is used to measure achievement of the four ARCS items above. The IMMS model scale consists of 36 items corresponding to attention (12 items), relevance (9 items), confidence (9 items), and satisfaction (6 items) of learning materials, which are rated on a five-point scale from “1. not at all applicable” to “5. completely applicable”.

2.3 Problem

For an earlier study, disaster prevention teaching system materials were constructed to conduct “evacuation training”, “firefighting training”, and “comprehensive training” while presenting a fire scene in a VR space to training participants. During evaluation of the ARCS model using the IMMS scale, the three items of “Attention,” “Relevance,” and “Satisfaction” were improved to the extent that significant differences from the existing

teaching materials were found. These results suggest that experiencing a disaster in a VR space improves these three items compared to the existing teaching materials. Only the item of “Confidence” was not improved to any great degree. One reason for this finding is the operability of VR.

For VR, users generally sit in front of a PC and use a controller to control an avatar on the VR screen to perform actions such as movement. In such cases, when participants try to perform actions such as evacuation while maintaining a low posture, as was done during the “evacuation drill”, it was necessary to convert “crouching” into “pressing the crouch button” and “moving” into “moving down the controller stick” simultaneously in the participants’ brains. Therefore, we infer that training participants perceive real-life actions and VR actions as different. This negative perception exerts a negative effect on their “Confidence” in their ability to perform the content of the training in reality. In addition, an important strength of VR training is that it can be repeated. The longer the training time in VR becomes, the greater the risk of motion sickness (VR sickness) to participants. Motion sickness causes severe nausea, dizziness, fatigue, and other symptoms.

3 Proposed System

3.1 Overview

Figure 1 presents a schematic diagram of the system materials proposed in this study.

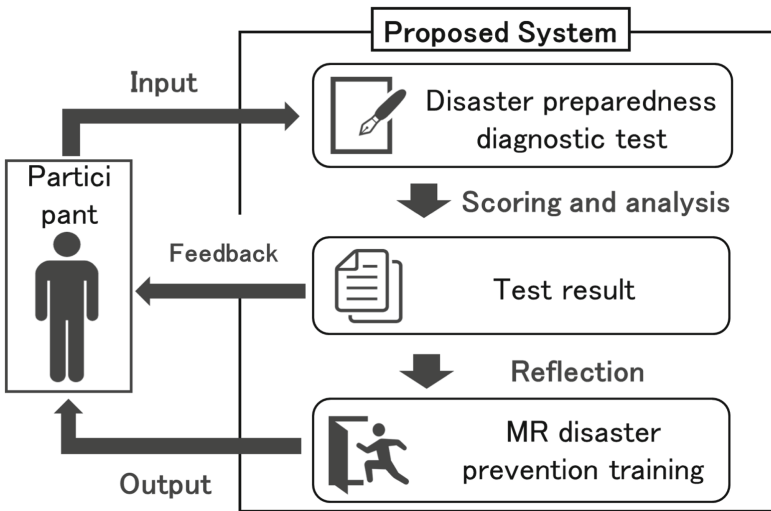


Fig. 1. Schematic diagram of the system materials proposed in this study.

First, we propose incorporation of a test of knowledge of disaster prevention of different fields into the system. Although individual differences exist, people do not think deeply or very often about disasters in daily life. Therefore, by providing each

participant with objective feedback for areas in which the participant lacks knowledge and skills in disaster prevention, the system is expected to provide an impetus for training in those areas and to make the participants feel the importance of the training.

Next, we propose a system for participants to experience and train for disasters in an MR space. The system has MR head-mounted displays (HMDs) such as HoloLens2 (Microsoft Corp.), using an environmental map of the real world and generating MR objects that can be manipulated similarly as in real space. In a real-space based space, one can realize training using the same actions as those used in real space.

Furthermore, because of its upgraded viewing angle and resolution, HoloLens2 can provide a high sense of presence and immersion to a user. Unlike VR, which completely replaces vision, HoloLens2 causes no discrepancy between the user’s motor and sensory systems. Therefore, it is expected to reduce motion sickness.

3.2 Implementing the Disaster Preparedness Diagnostic Test

The main purpose of incorporating the disaster preparedness diagnostic test into the system is identification of areas of disaster preparedness in which each participant lacks readiness. Finding areas in which each participant lacks preparedness can motivate them to train and compensate for a lack of preparedness. It will also help to create MR disaster preparedness scenarios to assess points of training that are lacking. Additionally, it might be possible to use this information to create scenarios for MR disaster drills to assess areas in which training is lacking.

For this study, we specifically examined two unpredictable disasters: earthquake and fire. We divided them into three areas of “initial response”, “evacuation response”, and “disaster prevention equipment, etc.”. For each, we created four questions for each of six areas (24 questions) to test the associated knowledge. Figure 2 presents an example of the classifications applied for the respective disasters.






Item	Initial response	Evacuation response	Disaster prevention equipment
Earthquake 	Shakeout 	Evacuation 	Fixing furniture 
Fire 	Awareness 	Evacuation 	Fire extinguisher 

Fig. 2. Examples of disaster categories.

Participants were tested on the iPython console with 25 questions, including 24 questions and one awareness question. They were given feedback based on the results.

3.3 Implementing MR Disaster Prevention Training

Implementation of the main part of this system, the MR disaster prevention training, is described hereinafter. The training progresses mainly through a blue window created by modifying the MRTK UI. The training participant moves to the next page by clicking the “Next” button in the window. To reduce the time spent gazing at the UI window, which is an MR object during training, voice guidance “CV: Sasara Sato (CeVIO Creative Studio)” was introduced. Introduction of voice guidance is expected to mitigate motion sickness by decreasing the time spent gazing at the UI window, which is an MR object, during the drill.

Four major scenes were created as the MR disaster drill contents. The training emphasized earthquake and fire disaster occurrence. They are particularly difficult to predict and wreak considerable destruction in Japan.

Scene 1, Menu screen: Screen for selecting content from Tutorial, MR Disaster Prevention Training (Earthquake), and MR Disaster Prevention Training (Fire).

Scene 2, Tutorial: A scene to familiarize the user with the UI window and MR object operations.

Scene 3.1, MR disaster drill (Earthquake version): A scene for earthquake experience in MR space and training for how to respond when an earthquake occurs.

Scene 3.2, MR disaster drill (Fire version): Scene in which participants experience a fire in MR space and practice how to respond to a fire.

Details of MR disaster drills (earthquake and fire) are presented below.

- **Scene3.1, MR disaster prevention training (Earthquake)**

The earthquake drill consists of 13 pages. It consists mainly of the following three topics: (1) initial response upon hearing an earthquake early warning (shakeout drill), (2) evacuation response when evacuated to an evacuation site because of an earthquake, and (3) earthquake preparedness.

Particularly, shakeout training should be conducted for each participant at various locations that they often encounter. The training is a good match with the HoloLens2 device, which can be activated as stand-alone.

The evacuation response and earthquake preparedness encourage participants to think about what they would do if they were in this situation and to write their report in a memo application on their smartphone.

This conjecture by participants is made possible by the fact that participants can refer to their smartphones in case of an emergency. It is also possible because HoloLens2 is a see-through HMD that allows them to operate devices other than HoloLens2 during training.

- **Scene 3.2, MR disaster drill (fire)**

Fire training, which also consists of 13 pages, consists mainly of the following three scenarios: (1) confirming the characteristics of fire and smoke, (2) evacuation by keeping

the upper body low, and (3) extinguishing a fire using a fire extinguisher, all of which are experienced in MR space.

Fire and smoke are set as a fire source by SLAM of HoloLens2 by searching for a likely fire source in the environment map creation and environment recognition. Then, during the evacuation drill, SLAM-based self-position estimation realizes a process by which the evacuation is completed when the user moves 5 m distant from the fire source. For firefighting using a fire extinguisher, MRTK hand-tracking is used to control the extinguisher. The Unity's game engine [5] functionality is used to ascertain when the extinguishing solution and flame mutually collide. Figures 3 and 4 portray the use of a fire extinguisher, an MR object, in MR space.

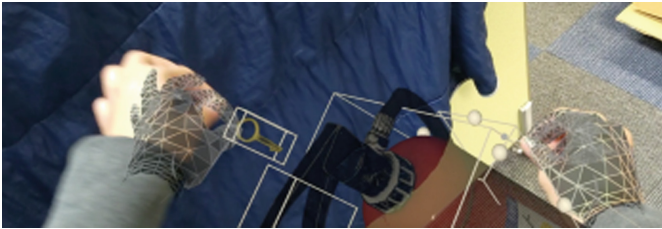


Fig. 3. Using a fire extinguisher: pulling out the safety pin.



Fig. 4. Using a fire extinguisher: point the hose at the fire.

4 Experimentation and Evaluation

4.1 Purpose of Experimentation

- To find improved evaluation values of “Confidence” after training

An earlier study was unable to confirm a significant difference in “Confidence” among the four items of the ARCS model in the VR teaching materials compared to

existing teaching materials. For this study, to confirm whether the evaluation value of “Confidence” was improved, we conduct a comparative evaluation experiment to assess the developed system.

- To investigate motion sickness risk

The VR teaching materials in the earlier study were regarded as entailing a risk of developing motion sickness (VR sickness). We investigate whether the system developed for this study entails a risk of motion sickness.

4.2 Experiment and Evaluation

To evaluate the system developed for this study, we conducted an evaluation experiment to assess the performance of six participants. The experiment was conducted according to the following procedure.

- Step 1: Preliminary interview survey to the motion sickness
- Step 2: Diagnostic test of disaster response capability
- Step 3: Disaster drill using MR disaster drill materials
- Step 4: Interview survey for onset of motion sickness
- Step 5: Questionnaire survey using the IMMS scale based on the ARCS model

A few days after the experiment, a questionnaire survey using the IMMS scale was conducted for “My Disaster Survival Notebook” [6], an existing teaching material, to compare evaluation values.

4.2.1 Comparison with Existing Teaching Materials Using the IMMS Scale

Table 1 presents the IMMS scale evaluation values for the ARCS model items for the system developed for this study.

Table 1. ARCS itemized assessment of proposed instructional material

Item	Average score of proposed material	Standard deviation
Attention	4.15	0.833
Relevance	3.91	1.051
Confidence	4.19	0.913
Satisfaction	4.06	0.984

Table 1 shows that all four ARCS items received high ratings overall, with a mean value of around 4 on a five-point scale. Next, a two-tailed *t*-test was conducted to evaluate the IMMS scale and its correspondence to the existing teaching material, “My Disaster Survival Handbook”. Table 2 presents the findings.

Table 2. Comparison of IMMS scale evaluation of existing and proposed material

Item	Average score of existing material	Average score of proposed material	Two-tailed <i>t</i> -test
Attention	2.65	4.15	**
Relevance	3.07	3.91	**
Confidence	2.98	4.19	**
Satisfaction	2.47	4.06	**

** : $p < 0.01$

Table 2 shows significant differences at the 1% significance level found for all four ARCS items, including the “Confidence” item, which was not found to be significantly different in an earlier study. Results indicate that all ARCS items of the developed disaster training material system were more motivating than existing materials.

Our consideration of results of IMMS scale comparison with existing materials is presented next. Significant difference was found for the “Confidence” item, which was not found in an earlier study. We inferred that the VR used for an earlier study showed a discrepancy between actions in training (controller-based operations) and actions in the real space, which might have affected improvement of the “Confidence” item. For this study, we used an MR space based on real space for training and eliminated the gap separating actions in the training and behaviors in the real space. As a result, an increase was observed for the “Confidence” item.

4.2.2 Survey of Motion Sickness Risk

The first question of the interview survey for motion sickness was “Have you ever experienced VR?” For participants who answered “Yes, I have experienced VR”, we asked whether they had ever experienced VR sickness when using VR. Then, irrespective of whether or not they had experienced VR, they were asked about whether they have awareness of being prone to screen sickness. Participants who were aware that they were prone to screen sickness were told to stop immediately if they felt any change in their physical condition. All participants were asked to proceed through the MR disaster drill material (about 40 min), after which they were interviewed to ascertain whether they felt any symptom of motion sickness.

The results are presented in Table 3.

The tendency to feel motion sickness varies from person to person. We discuss the tendencies of participants to feel motion sickness based on three pre-training questions.

First, Participant B had experienced VR and had experienced VR sickness a few times. The person was aware that she was prone to screen sickness. She was prone to motion sickness. In addition, although Participant D was not known to be disoriented by VR because he had never experienced VR, he was aware that he is prone to screen sickness. He was regarded as prone to motion sickness.

Table 3. Interview results and motion sickness of participants

	Subject A	Subject B	Subject C	Subject D	Subject E	Subject F
Experiencing VR	×	○	×	×	○	○
Experienced VR sickness	–	○	–	–	×	×
Prone to screen sickness	×	○	×	○	×	×
After a 40-min MR disaster drill						
Symptoms of motion sickness	×	×	×	×	×	×

Regarding Participants A and C, they were unaware that they were prone to screen sickness. Because they had never experienced VR, it was impossible to ascertain whether they were prone to motion sickness.

Finally, Participants E and F were unaware that they were prone to screen sickness, had experienced VR, and had not experienced VR sickness as a result of the experience. They are regarded as not being prone to motion sickness.

Because the present study was also a verification of motion sickness, we asked all six participants to complete the MR disaster drill (approximately 40 min long). We also interviewed them after the drill to ask whether they felt any motion sickness symptom, or not. None of the six participants reported any motion sickness symptom.

Our consideration of motion sickness risk survey results is presented next. The entire MR drill lasted approximately 40 min, which is sufficient time to ascertain motion sickness risk. No motion sickness symptom was observed for any of the six participants, including Participants B and D, who were regarded as particularly prone to motion sickness, suggesting that motion sickness risks of this system are low.

The MR space used for this study was based on real space. The participants' motor and sensory systems diverged only slightly, suggesting a suppressed risk of motion sickness. Furthermore, the addition of voice guidance to training progression might have reduced the amount of time spent gazing at the UI window during training.

5 Conclusion

This study proposed and examined disaster prevention training system materials by which participants' knowledge of disaster prevention was evaluated objectively before training. Based on the test results, training was conducted in the MR space.

The existing educational materials and the educational material system developed for this study were compared using the IMMS scale based on the ARCS model. Results confirmed significant differences for all ARCS items, including the "Confidence" item, which was not found to be significantly different in an earlier study. The educational materials developed for this study were demonstrated to be more motivating for learning disaster prevention than existing educational materials.

The risk of motion sickness was also confirmed to be sufficiently low with this system. None of the six participants, including two regarded as prone to motion sickness, reported motion sickness during the 40 min MR disaster drill.

Future studies will be conducted to brush up overall MR training materials to improve the system quality.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number JP20K11952.

References

1. Ministry of Land, Infrastructure, Transport and Tourism: White Paper on Land, Infrastructure, Transport and Tourism in Japan (2020)
2. <https://www.mlit.go.jp/hakusyo/mlit/r01/hakusho/r02/html/n1222000.html>
3. Tachi, S., et al.: Virtual Reality Science. The Virtual Reality Society of Japan (2011)
4. Milgram, P., Kishino, F.: A taxonomy of mixed reality visual displays. *IEICE Trans. Inf. Syst.* **E77-D**(12), 1321–1329 (1994)
5. Taisuke, T., Sano, M.: VR Disaster Training System for Improving Disaster Awareness. *IPSJ Interaction* (2018)
6. Unity. <https://unity.com/ja>
7. Fire and Disaster Management Agency. My Disaster Survival Notebook
8. <https://www.fdma.go.jp/relocation/syobodan/activity/education/bousai/survival/>

Author Index

A

Akita, Kazuki, 385
Ampririt, Phudit, 15, 27
Asada, Sora, 38

B

Baba, Akira, 125
Barolli, Admir, 1, 15, 159
Barolli, Leonard, 1, 15, 27, 38, 159, 167, 301, 323, 341, 353, 361
Batzorig, Munkhdelgerekh, 137, 215
Bouras, Christos, 45
Bylykbashi, Kevin, 1, 15, 27

C

Cante, Luigi Colucci, 248
Chatzigeorgiou, Charalampos, 45
Chen, Lung-Pin, 292
Chen, Tzu-Chi, 270
Chu, Yen-Chu, 282

D

Di Martino, Beniamino, 248
Duolikun, Dilawaer, 66, 180
Duulga, Baasantogtokh, 215

E

Enokido, Tomoya, 66, 78, 180
Esposito, Antonio, 248

F

Fan, Yao-Chung, 270
Fujioka, Hiroyuki, 377
Fujioka, Kaoru, 369

G

Graziano, Mariangela, 248
Guo, Shih-Wei, 270

H

Hashimoto, Hikari, 192
Hirata, Aoto, 38, 167, 353
Hirota, Masaharu, 361
Honda, Junichi, 312
Hoxha, Rexhina, 204

I

Ikeda, Makoto, 27, 301
Ikenaga, Takeshi, 224

K

Kakubari, Yasuyuki, 312
Katayama, Kengo, 323
Khwairakpam, Sanjukta, 236
Kim, Byungchan, 147
Kim, Yoonji, 137
Koga, Shiori, 369
Koh, Yeji, 137
Kollia, Anastasia, 45
Kulla, Elis, 15, 159, 341
Kuo, Chia-Chen, 292

L

Lai, Sen-Tarng, 259
Lee, Myung, 224
Leu, Fang-Yie, 259, 270, 282, 292

Li, Jung-Bin, 282
 Lin, Szu-Yin, 282
 Lin, Tzu-Ching, 292
 Liu, Yi, 159

M

Maeda, Hiroshi, 332
 Mandri, Eva, 204
 Matsunaga, Keisuke, 312
 Matsuo, Keita, 27, 341
 Meçe, Elinda Kajo, 204
 Mezawa, Yuki, 114
 Mimura, Mamoru, 114
 Miyake, Yutaka, 125

N

Nagai, Yuki, 38, 167, 353
 Nakahara, Masataka, 125
 Nakamura, Shigenari, 66, 78
 Niihara, Masahiro, 361
 Nishigaki, Masakatsu, 102, 125
 Nitta, Naoya, 125
 Nobayashi, Daiki, 224

O

Oda, Tetsuya, 38, 167, 323, 353, 361
 Ogiela, Lidia, 175
 Ogiela, Urszula, 175
 Oh, Insu, 215
 Ohki, Tetsushi, 102, 125
 Okudera, Ryosuke, 125
 Otsuyama, Takuya, 312
 Ouchi, Yumo, 125

P

Park, Hyunhee, 147
 Pouyioutas, Philippos, 45

Q

Qafzezi, Ermioni, 1, 15, 27

S

Sacile, Roberto, 204
 Saito, Nobuki, 38, 167, 323, 353, 361
 Sakamoto, Shinji, 1, 159
 Sekiguchi, Naoya, 91
 Serizawa, Ayumi, 125
 Shibata, Masahiro, 236
 Shigeyasu, Tetsuya, 192
 Shinko, Ilir, 204
 Shiomi, Yuya, 125
 Sinorukaj, Artemisa, 204
 Spaho, Evjola, 59

T

Tabuchi, Kei, 167
 Takata, Risa, 369
 Takeuchi, Ren, 102
 Takizawa, Makoto, 1, 66, 78, 159, 175, 180
 Tanaka, Hidema, 91
 Toyoshima, Kyohei, 38, 167, 323, 353, 361
 Tsukamoto, Kazuya, 224
 Tsuneyoshi, Mitsuki, 301
 Tsuru, Masato, 236

U

Uchiya, Takahiro, 385

W

Wang, Hui-Juan, 270
 Wang, Ming-Jen, 292

Y

Yagyū, Kohei, 102
 Yamashita, Yuma, 323
 Yano, Jinya, 377
 Yim, Kangbin, 137, 215
 Yoshihira, Mizuho, 125
 Yukawa, Chihiro, 323, 353, 361

Z

Zero, Enrico, 204