

James J. Park
Simon James Fong
Yi Pan
Yunsick Sung *Editors*

Advances in Computer Science and Ubiquitous Computing

CSA-CUTE 2019

Lecture Notes in Electrical Engineering

Volume 715

Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy

Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico

Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India

Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany

Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China

Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany

Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China

Gianluigi Ferrari, Università di Parma, Parma, Italy

Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain

Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany

Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

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

Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt

Torsten Kroeger, Stanford University, Stanford, CA, USA

Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA

Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore

Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany

Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA

Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany

Subhas Mukhopadhyay, School of Engineering & Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand

Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA

Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy

Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Gan Woon Seng, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Joachim Speidel, Institut of Telecommunications, Universität Stuttgart, Stuttgart, Germany

Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal

Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China

Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Associate Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Executive Editor (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

** Indexing: Indexed by Scopus. **

More information about this series at <http://www.springer.com/series/7818>

James J. Park · Simon James Fong · Yi Pan ·
Yunsick Sung
Editors

Advances in Computer Science and Ubiquitous Computing

CSA-CUTE 2019

 Springer

Editors

James J. Park
Department of Computer Science
and Engineering
Seoul University of Science and
Technology
Nowon-gu, Korea (Republic of)

Yi Pan
Department of Computer Science
Georgia State University
Atlanta, GA, USA

Simon James Fong
Faculty of Science
Department of Computer and Information
Science
University of Macau
Taipa, Macao

Yunsick Sung 
Department of Multimedia Engineering
Dongguk University
Seoul, Korea (Republic of)

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-15-9342-0

ISBN 978-981-15-9343-7 (eBook)

<https://doi.org/10.1007/978-981-15-9343-7>

© Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are reserved 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 Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Message from the CSA 2019 General Chair

International Conference on Computer Science and its Applications (CSA 2019) is the 11th event of the series of International Scientific Conference. This conference takes place in Macao, China, December 18–20, 2019. CSA 2019 will be the most comprehensive conference focused on the various aspects of advances in computer science and its applications. CSA 2019 will provide an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of CSA. In addition, the conference will publish high-quality papers which are closely related to the various theories and practical applications in CSA. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in this important subject. CSA 2019 is the next event in a series of the highly successful International Conference on Computer Science and its Applications, previously held as CSA 2018 (10th Edition: Kuala Lumpur, Malaysia), CSA 2017 (9th Edition: Taichung, Taiwan), CSA 2016 (8th Edition: Bangkok, Thailand, 2016), CSA 2015 (7th Edition: Cebu, December 2015), CSA 2014 (6th Edition: Guam, December 2014), CSA 2013 (5th Edition: Da Nang, December 2013), CSA 2012 (4th Edition: Jeju, November 2012), CSA 2011 (3rd Edition: Jeju, December 2011), CSA 2009 (2nd Edition: Jeju, December 2009), and CSA 2008 (1st Edition: Australia, October 2008).

The papers included in the proceedings cover the following topics: Mobile and ubiquitous computing, Dependable, reliable and autonomic computing, Security and trust management, Multimedia systems and services, Networking and communications, Database and data mining, Game and software engineering, Grid and scalable computing, Embedded system and software, Artificial intelligence, Distributed and parallel algorithms, Web and Internet computing, and IT policy and business management.

Accepted and presented papers highlight new trends and challenges of Computer Science and its Applications. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research. We would like to express our sincere thanks to Steering Chairs: James J. (Jong Hyuk) Park (SeoulTech, Korea), Yi Pan (Georgia

State University, USA), Han-Chieh Chao (National Ilan University, Taiwan), Young-Sik Jeong (Dongguk University, Korea), and Vincenzo Loia (University of Salerno, Italy).

Our special thanks go to the Program Chairs, Arun Kumar Sangaiah (VIT University, India), Mu-Yen Chen (National Taichung University of Science and Technology, Taiwan), Houcine Hassan (Universitat Politècnica de València, Spain); all Program Committee members; and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

CSA 2019 General Chair

Simon James Fong, University of Macau, Macau, China

Kyungeun Cho, Dongguk University, Seoul, Korea

Kim-Kwang Raymond Choo, The University of Texas at San Antonio, USA

Victor Leung, The University of British Columbia, Canada

Jungho Kang, Baewha Women's University, Korea

Message from the CSA 2019 Program Chairs

Welcome to the 10th International Conference on Computer Science and its Applications (CSA 2019) which will be held in Macao, China, December 18–20, 2019. CSA 2019 will be the most comprehensive conference focused on the various aspects of advances in computer science and its applications.

CSA 2019 provides an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of Computer Science. In addition, the conference contains high-quality papers which are closely related to the various theories and practical applications in Computer Science. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in this important subject. CSA 2019 is the next event in a series of highly successful International Conference on Computer Science and its Applications, previously held as CSA 2018 (10th Edition: Kuala Lumpur, Malaysia), CSA 2017 (9th Edition: Taichung, Taiwan), CSA 2016 (8th Edition: Bangkok, Thailand, 2016), CSA 2015 (7th Edition: Cebu, December 2015), CSA 2014 (6th Edition: Guam, December 2014), CSA 2013 (5th Edition: Da Nang, December 2013), CSA 2012 (4th Edition: Jeju, November 2012), CSA 2011 (3rd Edition: Jeju, December 2011), CSA 2009 (2nd Edition: Jeju, December 2009), and CSA 2008 (1st Edition: Australia, October 2008).

CSA 2019 contains high-quality research papers submitted by researchers from all over the world. Each submitted paper was peer-reviewed by reviewers who are experts in the subject area of the paper. Based on the review results, the Program Committee accepted papers.

For organizing an International Conference, the support and help of many people is needed. First, we would like to thank all authors for submitting their papers. We also appreciate the support from program committee members and reviewers who carried out the most difficult work of carefully evaluating the submitted papers.

We would like to give our special thanks to Profs. James J. (Jong Hyuk) Park, Yi Pan, Han-Chieh Chao, Young-Sik Jeong, and Vincenzo Loia, the Steering Committee Chairs of CSA, for their strong encouragement and guidance to organize the symposium. We would like to thank CSA 2018 General Chairs: Profs. Kim-Kwang Raymond Choo, Victor Leung, and Jungho Kang. We would like to express special thanks to committee members for their timely unlimited support.

CSA 2019 Program Chairs

Arun Kumar Sangaiah, VIT University, India

Mu-Yen Chen, National Taichung University of Science and Technology, Taiwan

Houcine Hassan, Universitat Politècnica de València, Spain

CSA 2019 Organization

Honorary Chair

Doo-Soon Park, Soonchunhyang University, Korea

Steering Committee

James J. Park, SeoulTech, Korea (Leading Chair)

Yi Pan, Georgia State University, USA

Han-Chieh Chao, National Ilan University, Taiwan

Young-Sik Jeong, Dongguk University, Korea

Vincenzo Loia, University of Salerno, Italy

General Chairs

Simon James Fong, University of Macau, Macau, China

Kyungeun Cho, Dongguk University, Seoul, Korea

Kim-Kwang Raymond Choo, The University of Texas at San Antonio, USA

Victor Leung, The University of British Columbia, Canada

Jungho Kang, Baewha Women's University, Korea

Program Chairs

Arun Kumar Sangaiah, VIT University, India

Mu-Yen Chen, National Taichung University of Science and Technology, Taiwan

Houcine Hassan, Universitat Politècnica de València, Spain

International Advisory Committee

Mo-Yuen Chow, North Carolina State University, USA
 Shu-Ching Chen, Florida International University, USA
 Mohammad S. Obaidat, Monmouth University, USA
 Enrique Herrera-Viedma, University of Granada, Spain
 Sherali Zeadally, University of Kentucky, USA
 Jordi Mongay Batalla, National Institute of Telecommunications, Poland
 Wanlei Zhou, Deakin University, Australia
 Sethuraman Panchanathan, Arizona State University, USA
 Yueh-Min Huang, National Cheng Kung University, Taiwan

Publicity Chairs

Ching-Hsien Hsu, Chung Hua University, Taiwan
 Ka Lok Man, Xi'an Jiaotong-Liverpool University, China
 Wei Song, North China University of Technology, China
 Deok Gyu Lee, Seowon University, Korea
 Fei Hao, Shaanxi Normal University, China

Program Committee

Chia-Hung Yeh, National Sun Yat-sen University, Taiwan
 Debajyoti Mukhopadhyay, Balaji Institute of Telecom and Management, India
 El-Sayed El-Alfy, King Fahd University of Petroleum and Minerals, Saudi Arabia
 Kuei-Ping Shih, Tamkang University, Taiwan
 M. Dominguez Morales, University of Seville, Spain
 Qian Yu, University of Regina, Canada
 Ana Isabel Pereira, Polytechnic Institute of Braganca, Portugal
 Antonina Dattolo, University of Udine, Italy
 Metin Basarir, Sakarya University, Turkey
 Javed Muhammad, Cornell University, Ithaca, NY, USA
 Sungsook Kim, Sun Moon University, Korea
 Yue-Shan Chang, National Taipei University, Taipei
 Ahmed EL Oualkadi, Abdelmalek Essaadi University, Morocco
 Haiduke Sarafian, Pennsylvania State University, USA
 Hiroyuki Tomiyama, Nagoya University, Japan
 Jie Shen, University of Michigan, USA
 Jung Hanmin, KISTI, Korea
 Liu Chuan-Ming, National Taipei University of Technology, Taipei
 Qingyuan Bai, Fuzhou University, China
 Sarkar Mahasweta, San Diego State University, USA
 Valev Ventseslav, Bulgarian Academy of Sciences, Bulgaria

Valle Mario, Swiss National Supercomputing Centre, Switzerland

Wojciech Zabierowski, Technical University of Lodz, Poland

Yu Chang Wu, Chung Hua University, Taiwan

Yutaka Watanobe, University of Aizu, Japan

Message from the CUTE 2019 General Chairs

On behalf of the organizing committees, it is our pleasure to welcome you to the 14th International Conference on Ubiquitous Information Technologies and Applications (CUTE 2019), which will be held in Macao, China, during December 18–20, 2019.

This conference provides an international forum for the presentation and showcase of recent advances in various aspects of ubiquitous computing. It will reflect the state-of-the-art computational methods, involving theory, algorithm, numerical simulation, error and uncertainty analysis, and/or novel application of new processing techniques in engineering, science, and other disciplines related to ubiquitous computing.

The papers included in the proceedings cover the following topics: Ubiquitous Communication and Networking, Ubiquitous Software Technology, Ubiquitous Systems and Applications, and Ubiquitous Security, Privacy, and Trust. Accepted papers highlight new trends and challenges in the field of ubiquitous computing technologies. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to Steering Committees: James J. Park (SeoulTech, Korea), Doo-Soon Park (Soonchunhyang University, Korea), Young-Sik Jeong (Dongguk University, Korea), Hsiao-Hsi Wang (Providence University, Taiwan), Laurence T. Yang (St. Francis Xavier University, Canada), Hai Jin (Huazhong University of Science and Technology, China), Chan-Hyun Youn (KAIST, Korea), Jianhua Ma (Hosei University, Japan), Mingyi Guo (Shanghai Jiao Tong University, China), and Weijia Jia (City University of Hong Kong, Hong Kong). We would also like to express our cordial thanks to the Program Chairs and Program Committee members for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference.

Finally, we thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

CUTE 2019 General Chairs

Simon James Fong, University of Macau, Macau, China

Yi Pan, Georgia State University, USA

Luis Javier Garcia Villalba, Universidad Complutense de Madrid, Spain

Message from the CUTE 2019 Program Chairs

Welcome to the 14th International Conference on Ubiquitous Information Technologies and Applications (CUTE 2019), which will be held in Macao, China, during December 18–20, 2019.

The purpose of the CUTE 2019 conference is to promote discussion and interaction among academics, researchers, and professionals in the field of ubiquitous computing technologies. This year the value, breadth, and depth of the CUTE 2019 conference continue to strengthen and grow in importance for both the academic and industrial communities. This strength is evidenced this year by having the highest number of submissions made to the conference.

For CUTE 2019, we received a lot of paper submissions from various countries. Out of these, after a rigorous peer-review process, we accepted only high-quality papers for the CUTE 2019 proceeding, published by Springer. All submitted papers have undergone blind reviews by at least two reviewers from the technical program committee, which consists of leading researchers around the globe. Without their hard work, achieving such a high-quality proceeding would not have been possible. We take this opportunity to thank them for their great support and cooperation.

Finally, we would like to thank all of you for your participation in our conference, and also thank all the authors, reviewers, and organizing committee members. Thank you and enjoy the conference!

CUTE 2019 Program Chairs

Muhammad Khurram Khan, King Saud University, Kingdom of Saudi Arabia

Neil Y. Yen, University of Aizu, Japan

Yunsick Sung, Dongguk University, Korea

CUTE 2019 Organization

Honorary Chair

Sanghoon Kim, Hankyong National University, Korea

Steering Committee

James J. Park, SeoulTech, Korea (Leading Chair)
Doo-Soon Park, Soonchunhyang University, Korea (Co-Chair)
Young-Sik Jeong, Dongguk University, Korea (Co-Chair)
Hsiao-Hsi Wang, Providence University, Taiwan
Laurence T. Yang, St. Francis Xavier University, Canada
Hai Jin, Huazhong University of Science and Technology, China
Chan-Hyun Youn, KAIST, Korea
Jianhua Ma, Hosei University, Japan
Mingyi Guo, Shanghai Jiao Tong University, China
Weijia Jia, City University of Hong Kong, Hong Kong

General Chairs

Simon James Fong, University of Macau, Macau, China
Yi Pan, Georgia State University, USA
Luis Javier Garcia Villalba, Universidad Complutense de Madrid, Spain

Program Chairs

Muhammad Khurram Khan, King Saud University, Kingdom of Saudi Arabia
Neil Y. Yen, University of Aizu, Japan
Yunsick Sung, Dongguk University, Korea

International Advisory Committee

Witold Pedrycz, University of Alberta, Canada
Seok Cheon Park, Gachon University, Korea
C. S. Raghavendra, University of Southern California, USA
Im-Yeong Lee, Soonchunhyang University, Korea
HeonChang Yu, Korea University, Korea
Hai Jin, Huazhong University of Science and Technology, China
Nammee Moon, Hoseo University, Korea
Byeong-Seok Shin, Inha University, Korea
Dong-Ho Kim, Soongsil University, Korea
Shu-Ching Chen, Florida International University, USA
Keun Ho Ryu, Chungbuk National University, Korea
JaeKwang Lee, Hannam University, Korea
Victor Leung, The University of British Columbia, Canada
Yoo-jae Won, Chungnam National University, Korea
Yang Xiao, The University of Alabama, USA

Publicity Chairs

Byoungwook Kim, Dongguk University, Korea
Jin Wang, Changsha University of Science & Technology, China
Deok Gyu Lee, Seowon University, Korea
Hyun-Woo Kim, Baewha Women's University, Korea
Seokhong Min, MINDATA Ltd., Korea
Joon-Min Gil, Catholic University of Daegu, Korea
Sung Chul Yu, LG Hitachi Co. Ltd., Korea
Yu-Wei Chan, Providence University, Taiwan
Jaehwa Chung, Korea National Open University, Korea
Jinho Park, Soongsil University, Korea
Hang-Bae Chang, Chung-Ang University, Korea

Program Committee

Bo-Chao Cheng, National Chung Cheng University, Taiwan
Chang Yao-Chung, National Taitung University, Taiwan
Dumitru Roman, SINTEF/University of Oslo, Norway
Eunmi Choi, Kookmin University, Korea
Heonchang Yu, Korea University, Korea
Imad Saleh, University of Paris 8, France
Jong-Myon Kim, University of Ulsan, Korea
Kwang Sik Chung, Korea National Open University, Korea
Yang-Sae Moon, Kangwon National University, Korea

Ali Shahrabi, Glasgow Caledonian University, UK
Bhekisipho Twala, University of Johannesburg, South Africa
Chen Uei-Ren, Hsiuping University of Science and Technology
Dugki Min, Konkuk University, Korea
Huang Kuo-Chan, National Taichung University of Education, Taiwan
HwaMin Lee, Soonchunhyang University, Korea
Jeong-Yong Byun, Dongguk University, Korea
Joon-Min Gil, Catholic University of Daegu, Korea
JungMin Kim, Daejin University, Korea
Kwangman Ko, Sangji University, Korea
Lai Kuan-Chu, National Taichung University, Taiwan
Lam-for Kwok, City University of Hong Kong, Hong Kong
Liang Tyng-Yeu, National Kaohsiung University of Applied Sciences, Taiwan
Omaima Bamasak, King Abdulaziz University, Saudi Arabia
Pinaki A Ghosh, Atmiya Institute of Technology and Science, India
Pyung-Soo Kim, Korea Polytechnic University, Korea
Serge Chaumette, University of Bordeaux 1, France
Seung Hyun Oh, Dongguk University, Korea
Stefanos Gritzalis, University of the Aegean, Greece
Wanquan Liu, Curtin University, Australia
Wookey Lee, Inha University, Korea

Contents

A Method for Nocturia Monitoring in Smart Home Using Decision Trees	1
Siriporn Pattamaset, Kwang Yong Kim, Min Kyu Joo, and Jae Sung Choi	
Efficient Data Aggregation for Human Activity Detection with Smart Home Sensor Network Using K-Means Clustering Algorithm	9
Siriporn Pattamaset and Jae Sung Choi	
Indoor Positioning System Using Pyramidal Beacon in Mobile Augmented Reality	17
Hyeon-woo An and Nammee Moon	
Design and Implementation of Real-Time Vehicle Recognition and Detection System Based on YOLO	25
Hyeonmoo Jeon, Gilwoo Lee, Byeongcheol Jeong, Jae Sung Choi, Jeong-Dong Kim, and Bongjae Kim	
Purchase Predictive Design Using Skeleton Model and Purchase Record	31
Jae-hyeon Cho and Nammee Moon	
Intelligent Digital Signage Using Deep Learning Based Recommendation System in Edge Environment	37
Kihoon Lee and Nammee Moon	
Performance Analysis of Single-Pulse Modulation in Factory Environment Based on LiFi Standard	45
Ho Kyung Yu and Jeong Gon Kim	
Deep Learning-Based Experimentation for Predicting Secondary Structure of Amino Acid Sequence	51
Syntia Widayuningtias Putri Listio, Ermal Elbasani, Tae-Jin Oh, Bongjae Kim, and Jeong-Dong Kim	

A Study on Vulnerabilities of Linux Password and Countermeasures 61
Sanghun Kim and Taenam Cho

Analysis of Learning Model for Improvement of Software Education in Korea 69
Ji-Hoon Seo and Kil-Hong Joo

Implementation and Experiment of Join Optimization Algorithm for Inverted Index in an RDBMS 79
Yoonmi Shin, Odsuren Temuujin, Minhyuk Jeon, Jinhyun Ahn, and Dong-Hyuk Im

Real-Time Subscriber Session Management on 5G NSA Wireless Network Systems 87
Kwan Young Park and Onur Soyer

PCA and K-means Based Genome Analysis for Hymenobacter sp. PAMC26628 93
Ermal Elbasani, So-Ra Han, Tae-Jin Oh, Bongjae Kim, and Jeong-Dong Kim

On Invariance of Concept Stability for Attribute Reduction in Concept Lattice 101
Fei Hao, Erhe Yang, Lantian Guo, Aziz Nasridinov, and Doo-Soon Park

A Study on Evidences Stored in Android Smartphones 107
Moses Kwak, Jisun Kim, Sungwon Lee, and Taenam Cho

Divide the FCA Network Graph into the Various Community Based on the *k*-Clique Methods 113
Phonexay Vilakone, Min-Pyo Hong, and Doo-Soon Park

An Efficient Micro-Service Placement Scheme Based on Fuzzy System for Edge-Enabled Digital Signage Service 121
A.-Young Son, Yeon Soo Lim, and Eui-Nam Huh

Rethinking Blockchain and Decentralized Learning: Position Paper 127
Sandi Rahmadika and Kyung-Hyune Rhee

A Study for Accelerating of Convolution Operations Based on Multiple GPUs with MPI 135
Boseon Hong, Geunmo Kim, Sungmin Kim, Jeong-Dong Kim, and Bongjae Kim

Requirements of Future Network for Blockchain Platform Operation 143
Suyeon Kim

TELL ME: Design of an Intelligence-Empowered Recommendation System 151
Kuan-Hua Lai, Neil Y. Yen, and Jason C. Hung

Estimation of Weights in Growth Stages of Onions Using Statistical Regression Models and Deep Learning Algorithm 159
 Wanhyun Cho, Junki Kim, Myung-Hwan Na, Sangkyoon Kim, and Hyejin Lee

Dynamic Projection Mapping Based on the Performer’s Silhouette 167
 Injae Jo, Youjin Koh, Taewon Kim, Sang-Joon Kim, Gooman Park, and Yoo-Joo Choi

Reinforcement Learning for Rate Adaptation in CSMA/CA Wireless Networks 175
 Soohyun Cho

Biometric-Based Seed Extraction Scheme for Multi-quadratic-Based Post-quantum Computing 183
 Aeyoung Kim and Seung-Hyun Seo

Lighting System to Maintain Color Temperature of Natural Light by Reflecting Changes of the Incoming Light 191
 Se-Hyun Lee, Seung-Taek Oh, and Jae-Hyun Lim

Deep Neural Network Model for Calculating Ultraviolet Information with Seasonal Characteristics from Illuminance 197
 Deog-Hyeon Ga, Dae-Hwan Park, Seung-Taek Oh, Heon-Tag Kong, and Jae-Hyun Lim

Model for Classifying Color Temperature Anomalies of Natural Light in Real Time Using Deep Learning 205
 Geon-Woo Jeon, Seung-Taek Oh, Heon-Tag Kong, and Jae-Hyun Lim

Low-Resolution LiDAR Upsampling Using Weighted Median Filter 213
 Hyun-bin Lim, Eung-su Kim, Pathum Rathnayaka, and Soon-Yong Park

Design of Tablet-Based Live Mobile Learning System Supporting Improved Annotation 221
 Jang Ho Lee

Gearbox Fault Diagnosis Under Variable Speed Condition Using Frequency Spectral Analysis with 1D Residual Neural Network 227
 Md Arafat Habib and Jong-Myon Kim

Health State Classification of a Spherical Tank Using a Hybrid Bag of Features and k-Nearest Neighbor 235
 Md Junayed Hasan, Jaeyoung Kim, and Jong-Myon Kim

L-RDFDiversity: Distributed De-Identification for Large RDF Data with Spark 243
 Minhyuk Jeon, Odsuren Temuujin, Yoonmi Shin, Jinhyun Ahn, and Dong-Hyuk Im

Intelligent Personalized Transport Alert System with Edge Computing 251
Hyolin Choi, Jiwon Hong, and Yongik Yoon

Induction Motor Bearing Fault Diagnosis Using Statistical Time Domain Features and Hypertuning of Classifiers 259
Rafia Nishat Toma and Jong-Myon Kim

Crack Detection Using Fully Convolutional Network in Wall-Climbing Robot 267
Myeongsuk Pak and Sanghoon Kim

Performance Evaluation of AODV and AOMDV Routing Protocols Under Collaborative Blackhole and Wormhole Attacks 273
Tran Hoang Hai, Nguyen Dang Toi, and Eui-Nam Huh

Simulation and Analysis of RF Attacks on Wireless SCADA System 281
Sung-Won Lee, Ji-Hun Kim, and Jonghee Youn

How Will Blockchain Technology Affect the Future of the Internet? 289
Geun-Hyung Kim

An Implementation of DAQ System for a Smart Fish Farm: Based on a Semi Circulation Filtration System in S. Korea 295
Joo H. Jean, Na E. Lee, Yoon H. Lee, Jea M. Jang, Moon G. Joo, Byung H. Yoo, and Jea D. Yoo

Design of Middleware to Support Auto-scaling in Docker-Based Multi Host Environment 301
Minsu Chae, Sangwook Han, and Hwa Min Lee

Gated Convolutional Neural Networks for Text Classification 309
Jin Sun, Rize Jin, Xiaohan Ma, Joon-young Park, Kyung-ah Sohn, and Tae-sun Chung

Algorithm Research of Face Recognition System Based on Haar 317
Xiaoguang Deng, Zijiang Zhu, Jing Chang, and Xiaojing Ding

Personal Authentication Based on EEG Signal and Deep Learning 325
Gi-Chul Yang

Security Information and Event Management Model Based on Defense-in-Depth Strategy for Vital Digital Assets in Nuclear Facilities 331
Sangwoo Kim, Seung-min Kim, Ki-haeng Nam, Seonuk Kim, and Kook-huei Kwon

A Web Archiving Method for Preserving Content Integrity by Using Blockchain 341
Hyun Cheon Hwang, Ji Su Park, Byung Rae Lee, and Jin Gon Shon

An SDN-Based Distributed Identifier Locator Separation Scheme for IoT Networks 349
Chan-Haeng Lee, Ji Su Park, and Jin Gon Shon

An Efficient Disposition for Wrist-Worn Device Usage Time Expansion in Wearable Computing Environment 357
Jong Won Lee, Ji Su Park, Hyeong Geun Kim, and Jin Gon Shon

Preserving Sustainability for Mission-Oriented Cyber-Physical Systems Collaboration 363
Horn Daneth, Nazakat Ali, and Jang-Eui Hong

A 3D Object Segmentation Method Using CCL Algorithm for LiDAR Point Cloud 371
Yifei Tian, Wei Song, Jinming Liu, and Simon James Fong

A Real-Time Human Posture Recognition System Using Internet of Things (IoT) Based on LoRa Wireless Network 379
Wei Song, Jinqiao Liao, and Jinkun Han

Mobile Charger Planning for Wireless Rechargeable Sensor Network Based on Ant Colony Optimization 387
Fan-Hsun Tseng, Hsin-Hung Cho, and Chin-Feng Lai

Deep Learning Based Malware Analysis 395
Sunoh Choi

CNN-GRU-Based Feature Extraction Model of Multivariate Time-Series Data for Regional Clustering 401
Jinah Kim and Nammee Moon

DNN-Based Mutual Satisfaction Prediction Model for Matching Between Users 407
Hyunnoh Yun, Jinah Kim, and Nammee Moon

An Approach to Improving Software Security Through Access Control for Data in Programs 413
Hyun-il Lim

Implementation of a Container-Based Interactive Environment for Big-Data Analysis on Supercomputer 421
Seungmin Lee, Ju-Won Park, Kimoon Jeong, and Jaegyeon Hahm

Inflight Tracking Method with Beacon System and Scouting Drone 427
Yunseok Chang

Deep Learning Based Character Recognition Platform in Complex Situations 435
BoSeon Kang, Seong-Soo Han, You-Boo Jeon, and Chang-Sung Jeong

Design of Restricted Coulomb Energy Neural Network Processor for Multi-modal Sensor Fusion 441
 Jaechan Cho, Minwoo Kim, Yongchul Jung, and Yunho Jung

Low Complexity Pipelined FFT Processor for Radar Applications 447
 Yongchul Jung, Jaechan Cho, and Yunho Jung

Dynamic Mitigation of Catastrophic Forgetting Using the Sampling Network 455
 Dae Yong Hong, Yan Li, and Byeong-Seok Shin

A Study on the Implementation of GRU Autoencoder Model for Detecting Insider Anomaly Behavior 461
 Kyeong Geun Ryu and Deok Gyu Lee

Blockchain Based Authentication Method for ThingsBoard 471
 Sung Il Jang, Ji Yong Kim, Alisher Iskakov, M. Fatih Demirci, Kok Seng Wong, Young Jong Kim, and Myung Ho Kim

Secure Management of Patient Medical Data Using QR Code and CP-ABE 481
 Su-Mee Moon, Beakcheol Jang, Hoon Yoo, and Jong Wook Kim

Restore Fingerprints Using Pix2Pix 489
 Ji-Hwan Moon, Jin-Ho Park, and Gye-Young Kim

Web Site Usage History Management System Using Blockchain 495
 Cheolmin Yeom, Seonghwa Yeon, Sunghyun Yu, and Yoojae Won

Generation of Fake Iris Images Using CycleGAN 503
 Jae-gab Choi, Jin-Ho Park, and Gye-Young Kim

Robust 3D Reconstruction Through Noise Reduction of Ultra-Fast Images 509
 Nu-lee Song, Jin-Ho Park, and Gye-Young Kim

Pedestrian Detection Using Regression-Based Feature Selection and Disparity Map 515
 Chung-Hee Lee

Blockchain-Based Multi-fogcloud Authentication System 521
 Jae Hwan Kwon, Young Kook Kim, Askhat Temir, Kamalkhan Artykbayev, M. Fatih Demirci, and Myung Ho Kim

Activity-Recognition Model for Violence Behavior Using LSTM 529
 Svetlana Kim, Hyejeong Nam, Hyunho Park, Yong-Tae Lee, and Yongik Yoon

Static Analysis for Malware Detection with Tensorflow and GPU 537
 Jueun Jeon, Juho Kim, Sunyong Jeon, Sungmin Lee, and Young-Sik Jeong

IoT Malware Dynamic Analysis Scheme Using the CNN Model 547
 Jueun Jeon, Seungyeon Baek, Minho Kim, Inho Go, and Young-Sik Jeong

A Design of Improvement Method of Central Patch Controlled Security Platform Using Blockchain 555
Kyoung-Tack Song, Shee-Ihn Kim, and Seung-Hee Kim

A Suggestion for ERP Software Customization Model Using Module Modification Factors 563
Byung-Keun Yoo and Seung-Hee Kim

A Location-Based Solution for Social Network Service and Android Marketing Using Augmented Reality 569
Jun-Ho Huh and Yeong-Seok Seo

Artificial Intelligence Based Electronic Healthcare Solution 575
Seong-Kyu Kim and Jun-Ho Huh

Optimal Location Recommendation System for Offshore Floating Wind Power Plant Using Big Data Analysis 583
Sang-Hyang Lee and Jun-Ho Huh

Efficient Data Noise-Reduction for Cyber Threat Intelligence System 591
Seonghyeon Gong and Changhoon Lee

An Improved DBSCAN Method Considering Non-spatial Similarity by Using Min-Hash 599
Jin Uk Yoon, Byoungwook Kim, and Joon-Min Gil

A Method for Nocturia Monitoring in Smart Home Using Decision Trees



Siriporn Pattamaset, Kwang Yong Kim, Min Kyu Joo, and Jae Sung Choi

Abstract Nocturia is widely regarded as urological with the symptom of the need to wake during bedtime one or more times, which causes to sleep loss. In this paper, we presented the method for nocturia monitoring with number of times of urination during bedtime and duration time of sleep before wake to urinate. The proposed method is the use of a few required sensors and decision trees. This method can be used for a follow-up to the symptom of a person with nocturia.

Keyword Sleeping monitoring · Activity detection · Urination · Micturition

1 Introduction

Nocturia or frequent nighttime urination defined as the necessity to wake to urinate during bedtime by the International Continence Society (ICS) [1]. People with nocturia does arouse from sleep to voluntarily urinate that differs from enuresis or bedwetting. The symptom of nocturia is the complaint that the individual has to wake at night one or more times to void [1].

This paper presents a method for nocturia monitoring in smart home using decision trees. The use of required sensors and tree models so as to produce nocturia monitoring report to a caregiver. The contribution of this proposed method can be used for follow-up to the symptom of a person with nocturia.

S. Pattamaset · K. Y. Kim · M. K. Joo · J. S. Choi (✉)
Department of Computer Engineering, Sun Moon University, Asan, Chungcheongnam-Do,
Republic of Korea
e-mail: jschoi@sunmoon.ac.kr

S. Pattamaset
e-mail: siriporn@sunmoon.ac.kr

K. Y. Kim
e-mail: dllrks97@sunmoon.ac.kr

M. K. Joo
e-mail: yioyio5623@sunmoon.ac.kr

2 Related Work

Nowadays, these few studies of nocturia monitoring. The first study developed a low-cost device that detects micturition event based on the impedance method [2]. In [3] presented an unobtrusive and non-stigmatizing device, based on an ambulatory, with measurement of bedtime micturition purpose. Another study described in [4], where a proof of concept application of the use of pressure mats to monitor nocturnal fluid intake and bladder voiding events can be tracked.

In this work, we propose the method to monitor nocturia through a count of times of urination during bedtime and duration time of sleep before wake to urinate using a few required sensors and decision trees.

3 Proposed Method

In this section, we will explain our Nocturia monitoring method which consists of five processes: measured data access, data pre-processing, sleeping detection, bedtime urination detection, and nocturia monitoring report. The step of the procedure is shown in Fig. 1. The data measuring from sensors are accessed and processed for missing value case. The other processes will be explained in detail in Sects. 3.1, 3.2, and 3.3.

3.1 Sleeping Detection Tree Model

Our method for sleeping detection employs decision trees, we build the tree model with the use of a few numbers of sensors with good performance of activity detection. The tree model is not only built for sleeping detection but also considering the case of person wake at night to urinate in order to achieve our purpose. The result of this process is the prediction of whether a person is sleeping or not.

We explore the main required sensors that smart home must have them. Although, more required sensors might be needed since a difference of environments and a person's behavioral pattern. Figure 2 shows the sleeping detection tree model with the main required sensors which are bedroom door sensor, bathroom door sensor, and bedroom luminosity sensor.

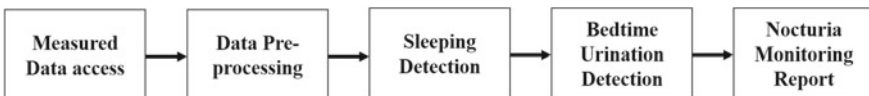


Fig. 1 The process of the proposed method

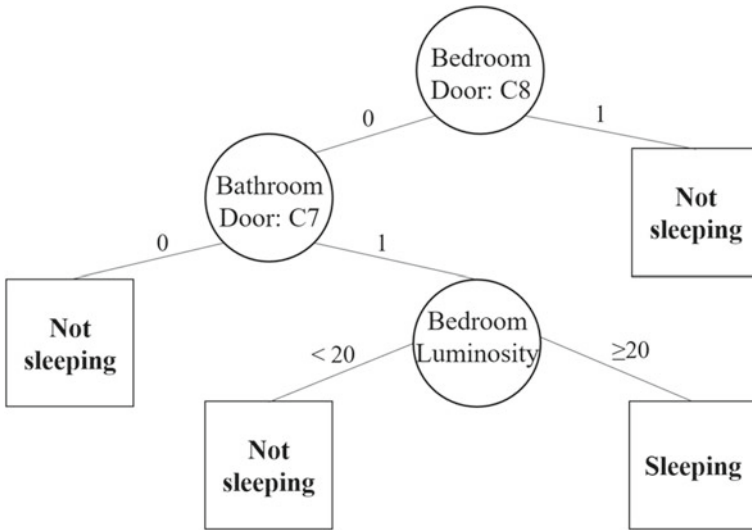


Fig. 2 The tree model for sleeping detection

3.2 Bedtime Urination Detection Tree Model

Bedtime urination detection tree model will be used for prediction in case of a person may wake at night to urinate. The model utilizes the same required door sensors and historical sleeping activity at 30 ms before. Even though this process is the prediction of a situation when the person likely wakes up to urinate during bedtime, the result is not a final result that will be sent to a caregiver since the final result will be analyzed and produced in Fig. 3.

3.3 Nocturia Monitoring Report

This process analyzes the output from the previous process to produce nocturia monitoring report to a caregiver. The report contains the information as follows: total sleep time, number of times that a person wakes to urinate during bedtime, and duration time of sleep before wake to urinate. The report was produced using an algorithm as shown in Table 1.

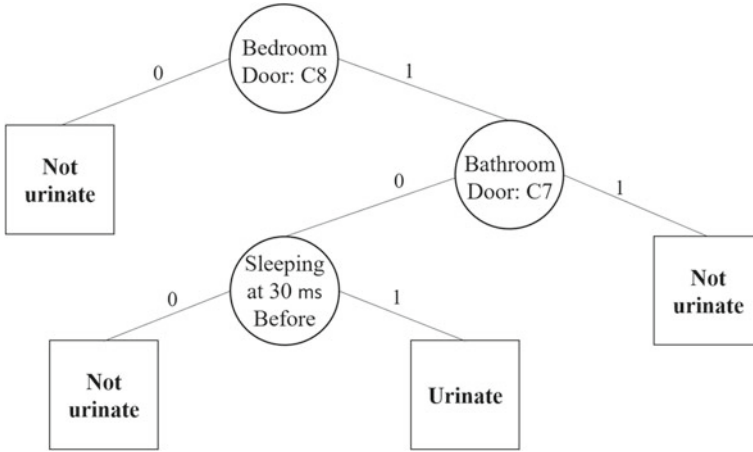


Fig. 3 The tree model for detecting when a person likely wakes to urinate during bedtime

4 Experiments and Results

The experiment was on a real-life smart home dataset, ContextAct@A4H Real-Life dataset [5]. Therefore, we can explore the main required sensor for sleeping detection with nocturia monitoring purpose. The dataset provides a dataset in two different periods, 7 days in July 2016 and 21 days in November 2016. An occupant annotated a start and stop time of activity. Nevertheless, the dataset noticed that the occupant forgot to annotation some activities.

Even the dataset was recorded from 28-year-old woman, who is not a person with nocturia, our method was proved that can detect urination during bedtime. The result of sleeping and urination during bedtime prediction on July dataset is shown in Fig. 4 and November dataset in Fig. 5, where the orange line stands for the prediction result from sleeping decision tree model, purple line stand for the prediction result from bedtime urination detection tree model, and purple line within dash line stands for the result from nocturia monitoring report process. In Fig. 6 and Fig. 7 showed the report that will be reported to a caregiver in July and November respectively. The July 26 and November 24 at 2nd urination reports were wrongly reported urination during bedtime, which is the effect of the wrongly sleeping detection result as shown in Fig. 5. The overall performance of our method is shown in Table 2.

5 Discussion and Conclusion

In this paper, we proposed a method for nocturia monitoring in smart home using decision trees which the use of a few of the required sensors. The results showed that our method was able to predict sleeping and urination during bedtime, in other

Table 1 The pseudocode for Nocturia detection report

Pseudocode Nocturia detection report

Input: PredictedSleeping(output from 3.1) as 1 or 0,
 PredictedBedtimeUrinate(output from 3.2) as 1 or 0

Output: report.txt

Initialization: StartSignWakeToUrinate = 0

```

1: if PredictedSleeping == 1
2:   SET SleepStart = CurrentTime
3: elseif PredictedSleeping == 0
4:   SET SleepEnd = CurrentTime
5: if PredictedSleeping == 1 and SleepEnd != null
6:   SET SleepStart2 = CurrentTime
7:   if SleepStart2 - SleepEnd < 1 hour
8:     Reset SleepStart2, SleepEnd
9:   else
10:    for SleepStart to SleepEnd
11:      if StartSignWakeToUrinate == 1
12:        if PredictedBedtimeUrinate == 0
13:          StartSignWakeToUrinate = 0
14:        else
15:          continue;
16:      if PredictedBedtimeUrinate == 0
17:        StartSignWakeToUrinate = 0
18:      elseif PredictedBedtimeUrinate == 1
19:        if getReal_Sleeping == 1
20:          continue;
21:        if PredictedSleeping == 0
22:          set StartSignWakeToUrinate =1
23:          if StartSignWakeToUrinate == 1 and
24:             PredictedBedtimeUrinate==1
25:             NumAnalyzedUrinate++
26:             Write result to text file
27:             SleepStart = SleepStart2
28:             Reset all parameters except SleepStart
29: return report

```

word, it can be used for nocturia monitoring since the system reports a number of times of urination during bedtime and duration time of sleep before wake to urinate, which can be used for follow-up the symptom of a person with nocturia.

The wrongly predicted results can see on the Figs. 4 and 5 that mostly seem to mistake from the cause of the dataset lacked annotation of some activities. Additionally, the mistake of sleeping prediction can affect urination during bedtime report.

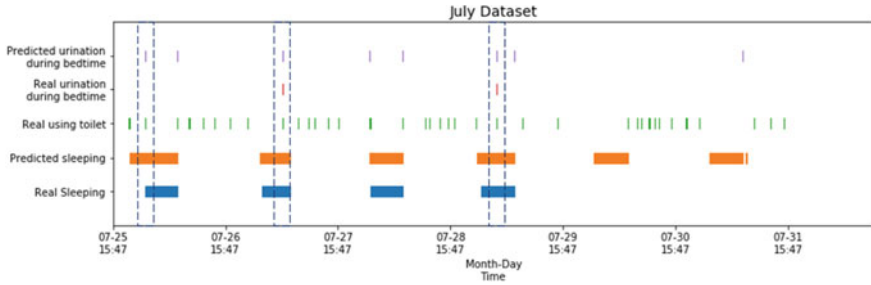


Fig. 4 The result of predicted sleeping and urination activity (during bedtime) compares to real activity for July dataset

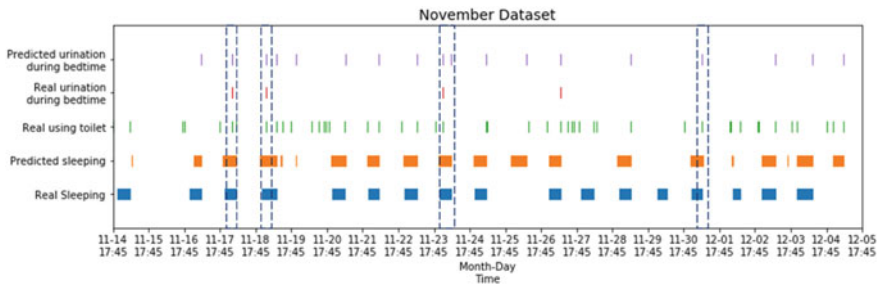


Fig. 5 The result of predicted sleeping and urination activity (during bedtime) compares to real activity for November dataset

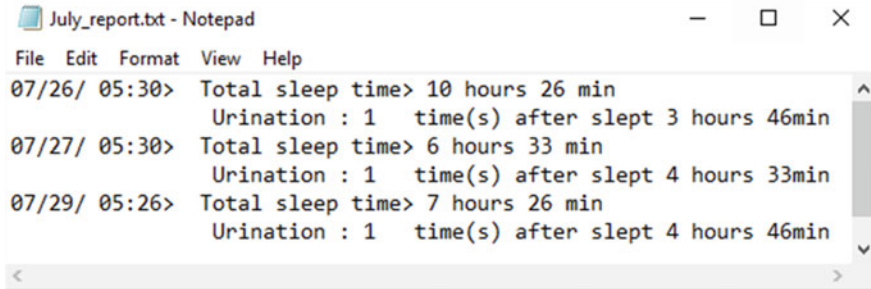


Fig. 6 The report of nocturia monitoring in July dataset

```

Nov_report.txt - Notepad
File Edit Format View Help
11/18/ 04:48> Total sleep time> 8 hours 26 min
                Urination : 1 time(s) after slept 5 hours 26min
11/19/ 07:47> Total sleep time> 11 hours 13 min
                Urination : 1 time(s) after slept 4 hours 26min
11/24/ 05:05> Total sleep time> 8 hours 33 min
                Urination : 1 time(s) after slept 2 hours 40min
                Urination : 1 time(s) after slept 5 hours 46min
12/01/ 06:41> Total sleep time> 8 hours 33 min
                Urination : 1 time(s) after slept 7 hours 46min
    
```

Fig. 7 The report of nocturia monitoring in November dataset

Table 2 The performance of sleeping activity prediction

Performance metrics	July dataset	November dataset	Average (%)
Accuracy	88.72	90.72	89.72
Recall	99.62	78.53	89.08
Precision	53.33	78.16	65.75
F1	69.47	78.34	73.91

Acknowledgements This work is result of studies on the “Leaders in INdustry-university Cooperation” Project, which is supported by the Korean Ministry of Education, and Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) 2018-0-01865, Sunmoon University, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5045953).

References

1. Abrams P, Cardozo L, Fall M, Griffiths D, Rosier P, Ulmsten U, Van Kerrebroeck P, Victor A, Wein A (2002) The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society. *Neurourol Urodyn Off J Int Cont Soc* 21:167–178
2. Taramasco C, Rodenas T, Martinez F, Fuentes P, Munoz R, Olivares R, Albuquerque VHC, Demongéot J (2018) A novel low-cost sensor prototype for nocturia monitoring in older people. 6:52500–52509
3. Huppert V, Paulus J, Paulsen U, Burkart M, Wullich B, Eskofier BM (2015) Quantification of nighttime micturition with an ambulatory sensor-based system. *IEEE J Biomed Health Inform* 20:865–872
4. Cohen-McFarlane M, Green JR, Knoefel F, Goubran R (2016) Smart monitoring of fluid intake and bladder voiding using pressure sensitive mats. In: 2016 38th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 4921–4924

5. Lago P, Lang F, Roncancio C, Jiménez-Guarín C, Mateescu R, Bonnefond N (2017) The ContextAct@ A4H real-life dataset of daily-living activities. In: International and interdisciplinary conference on modeling and using context. Springer, Cham, pp 175–188

Efficient Data Aggregation for Human Activity Detection with Smart Home Sensor Network Using K-Means Clustering Algorithm



Siriporn Pattamaset and Jae Sung Choi

Abstract Smart home sensor network utilizes various sensors to measure physical and send data to a base station. The pattern of measured data in each room can be considered as an active pattern when activity occurrence in that room and a irrelevant pattern when no activity in the room. In order to improve data aggregation in smart home, we propose human activity pattern-based data aggregation, which applies K-means clustering algorithm based on human activity into cluster heads of cluster-based sensor network. The result of simulation shows that the clustering algorithm can detect active event by calculating the similarity between the active pattern of collected data and human activity according to room usage.

Keywords Cluster head · Human activity · Pattern recognition · WSN

1 Introduction

Smart home operates over wire or wireless sensor network which needs some techniques to aggregate occupant activity through embedded sensor nodes for many purposes, such as assisted living, health monitoring, and operation of home appliances [1–3]. In order to thoroughly aggregate information in smart home, it might need to deploy a large number of ambient sensors. Indeed, the information consists of useful and useless information. However, data aggregation techniques help to collect useful information in a region of interest. As a consequence, energy consumption can be decreased and the network lifetime [4] can be increased for ensuring sustainable operation of smart home.

In this work, we design smart home sensor network with a cluster-based hierarchical architecture which consists of three levels as follows: sensor node, cluster head,

S. Pattamaset · J. S. Choi (✉)
Department of Computer Engineering, Sun Moon University, Asan, Chungcheongnam-Do,
Republic of Korea
e-mail: jschoi@sunmoon.ac.kr

S. Pattamaset
e-mail: siriporn@sunmoon.ac.kr

© Springer Nature Singapore Pte Ltd. 2021
J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_2

and base station. We focus on data aggregation in cluster heads of data representing human activity in smart home. In data aggregation, sensor nodes produce dataset providing real-world scenario parameters that can be clustered into different patterns of event. For examples, a pattern of human activity occurrence (active pattern) and a pattern of no activity occurrence (irrelevant event). The patterns of data measuring from sensor nodes can be clustered by using a clustering algorithm.

This paper presents an aggregation technique of using K-means clustering algorithm for clustering collected data patterns based on human activity in smart home sensor network. The collected data in cluster heads are grouped and selected the pattern that can represent human activity based on room usage. The contribution of this proposed approach is to reduce the size of data passing the transmission and energy consumption in a smart home. Additionally, it can be applied to use in human activity classification purpose as data pre-processing to be features of a classifier.

2 Related Work

The different ways are used to operate the internal communication network between sensor nodes and a base station and help to aggregate data that can improve the lifetime of the network. Maraiya et al. [5] categorized data aggregation approach into four approaches based on methodologies including Tree-based algorithm, Cluster-based algorithm, Multipath-based algorithm, and Hybrid-based algorithm.

Our sensor network is a cluster-based sensor network that can reduce the bandwidth overhead because of less of transmitted data packets since this approach has several cluster heads that gather data from numerous source nodes. Each cluster head needs intelligently aggregation data measured from numerous sensors in smart home for human activity interest and network lifetime by finding a pattern of human activity representation among whole data in a cluster head fixed and distributed based on room usage. The clustering algorithm is one of unsupervised learning algorithm that can be used to finding different patterns in a dataset without a label.

One of the simple clustering algorithms is K-means algorithm which considers mean of feature into k groups based on Euclidean distance. K-means algorithm has been used to enhance data aggregation in sensor networks. Harb et al. [6] used the K-means clustering algorithm group similar data sets into generated clusters before applying a prefix frequency filtering (PFF) in periodic sensor networks. They presented KPFF technique which enhanced a PFF technique using K-Means based clustering approach for data aggregation. The same researchers enhanced their similar previous purpose based on a one-way ANOVA model to identify generating nodes identical data sets and to aggregate these sets before sending them to the sink in [7] so as to eliminate redundancy from sensor node members that generate redundant data sets. Idrees et al. [8] used a modified K-means technique to remove the data redundancy in data aggregation for enhancement the lifetime of the sensor network. They proposed a Distributed Data Aggregation based Modified K-means (DiDAMoK) Technique, which consists of three stages. The first stage, a sensor node measures

and collects the data. The second stage, the modified K-means is employed on the collected data to convert them into clusters of them. The last stage is to transmit the representative collected data of each cluster to the sink.

Most of the researchers used the K-means algorithm to eliminate redundant data in the sensor network. Thanks to K-means algorithm provides the ability to group sensor information that can capture redundant data. In other words, K-means algorithm could be able to find pattern of human activity in collected data that can enhance data aggregation in smart home.

3 Human Activity Pattern-Based Data Aggregation

This section explains our data aggregation technique based on human activity, which utilizes correlation of measured information from sensor nodes and human activity in a smart home environment. Our approach enhances the performance of data aggregation by collecting data representing human activity and ignoring other data packets that are unrelated to human activity based on room usage.

In cluster-based sensor network, the physical environment and human activity are measured through all sensor nodes employed throughout the smart home. The sensor nodes updates data every minute (e.g., $s_{t,1}^r = 1$, $s_{t,2}^r = 25.5$, \dots , $s_{t,n}^r$), where n is the total number of data measured from sensor nodes in the cluster head r at t timestamp. The cluster head level consists of all cluster heads distributed into each room based on usage. Every cluster head plays a role of aggregator that receives data measured from their own children placed in the same room with them. The data are aggregated as CH_t^r vector (e.g., $CH_t^{bathroom} = \{s_{t,1}^{bathroom}, s_{t,2}^{bathroom}, \dots, s_{t,n}^{bathroom}\}$, $CH_t^r = \{s_{t,1}^r, s_{t,2}^r, \dots, s_{t,n}^r\}$), where r is the name of room. The CH_t^r vector contains much data measured from sensor n nodes at t timestamp.

The example of information in kitchen cluster head vector ($CH_{t=\{0,\dots,T\}}^{kitchen}$) includes both information that has a correlation to the occurrences of human activity and information that does not contain any relevant data or wanted data. In order to find pattern of collected data in each cluster head, we need to do offline learning from historical data before using for data aggregation in the cluster head. We use a clustering algorithm to cluster patterns of collected data at cluster head when it possibly contains data of activity occurrence so that the collected data is aggregated before forwarding to the base station. The clustering algorithm applies the simplest unsupervised learning, which is k-means++ algorithm [9]. The algorithm utilizes a whole data of cluster head in a dataset so as to cluster patterns of collected data into k clusters. We calculate an appropriate number of k (cluster) in each cluster head using Elbow method. The visualization in Fig. 1 shows the average of within-cluster sum of square distances and the optimal number of k at the red cycle.

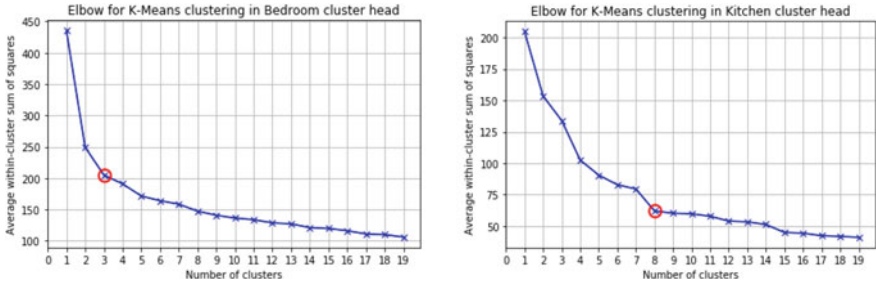


Fig. 1 The Elbow method plotted from bedroom (left) and kitchen (right) cluster heads

4 Simulation and Results

The simulation was performed to validate our proposed approach on ContextAct@A4H Real-Life dataset [10], which is the rich-sensor smart home scenario dataset. The dataset provides a dataset of daily living activities during July 2016 (1 week) and November 2016 (3 weeks) in an apartment equipped with 219 sensors and actuators. All sensors are ambient sensors deployed in a bedroom, bathroom, kitchen, study room, and around hallways. We used the November dataset with 168 data properties from different sensors to find data patterns of sleeping, cooking, taking shower, and using toilet activity. We set cluster heads based on room usage and gather data properties (bathroom (46), bedroom (57), kitchen (65)) from sensor nodes in each room. The dataset is log dataset with activities annotation (start time and stop time). However, the occupant reported missing some annotation of activity annotation. We modified the loc dataset to time series dataset with minute interval since we set all sensor nodes to transmit data to cluster head every minute.

The sensors are locally distributed based on room functioning and defined a fixed cluster head to each of them. Every cluster head applies the clustering algorithm to learn patterns of sensors while human activity is happening in a room (active event) and nothing is happening in the room or disinterested activity happening (irrelevant event). The example of sensor pattern clustering results is shown in Fig. 2. The number of patterns is following optimal k from Elbow method. Nevertheless, the correct number of optimal k is often ambiguous. Thus, we choose one cluster,

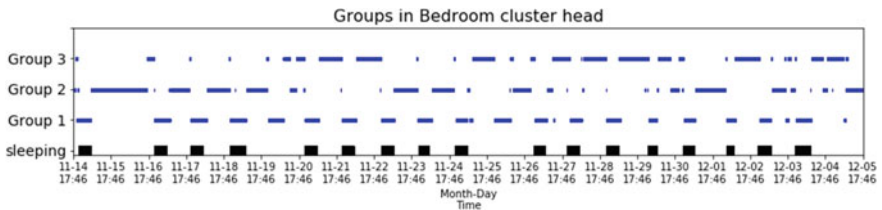


Fig. 2 The matching between different transmission time of bedroom cluster head and the activity

represents information when human activity occurrence (active event), among all clusters in different k from two clusters to k clusters by calculating Eq. (1), where M is a cluster that has a maximum of the percentage of the correspondence with information in each cluster and activity label, $G_{i,t}$ is a label of transmitting time at t in a cluster i , A_t is a relative activity label, with $G_{i,t}, A_t \in \{0, 1\}$.

$$M = \underset{i=\{1,\dots,k\}}{\operatorname{argmax}} \left(\frac{\sum_{t=0}^T (G_{i,t} \leftrightarrow A_t)}{T} \right) \tag{1}$$

The chosen cluster that represents when data packet is aggregated in a cluster head as shown in Fig. 3. The performance of our data aggregation technique is calculated from percentage of active event in each cluster head matching with the relative activity as shown in Table 1.

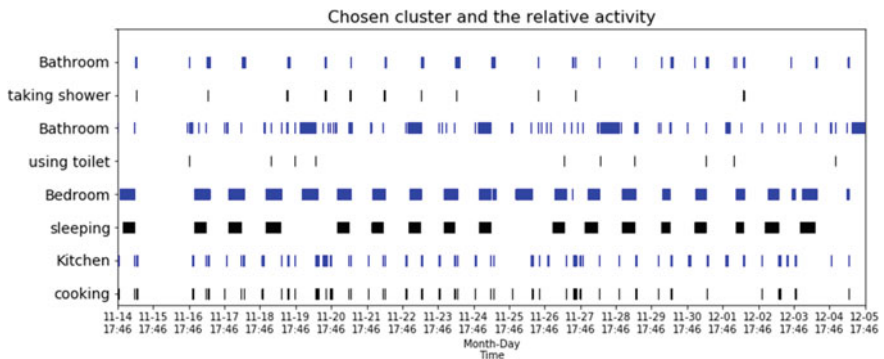


Fig. 3 The chosen clusters and the relative activity

Table 1 The percentage of active event in each cluster head matching with the relative activity

Cluster head	Chosen cluster i from k clusters	Relative activity	Matching (%)
Bathroom	$i = 5, k = 5$	Taking shower	96.2
Bathroom	$i = 1, k = 7$	Using toilet	88.5
Bedroom	$i = 1, k = 3$	Sleeping	90.4
Kitchen	$i = 1, k = 3$	Cooking	97.4

5 Discussion and Conclusion

This approach is able to improve data aggregation in order to contribute to less energy consumption in smart home for the purpose of human activity interest. Our approach uses of K-means clustering algorithm to find patterns of active event in cluster head. However, the human activity pattern-based data aggregation can be used when we know a relative activity in the room, such as kitchen, bedroom, and bathroom. On the contrary, other cluster heads that have no specific activity or multi-functional usage in the area, it will be difficult to cluster patterns of wanted data in the area.

In the future, we are going to apply human activity pattern-based data aggregation technique into smart home sensor network before transmitting collected data to a base station for human activity classification purpose so as to design sustainable sensor network of smart home.

Acknowledgements This work is result of studies on the “Leaders in INdustry-university Cooperation” Project, which is supported by the Korean Ministry of Education, and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) 2018-0-01865, Sunmoon University, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5045953).

References

1. Rashidi P, Mihailidis A (2013) A survey on ambient-assisted living tools for older adults. *IEEE J Biomed Health Inform* 17:579–590
2. Yassine A, Singh S, Alamri A (2017) Mining human activity patterns from smart home big data for health care applications. *IEEE Access* 5:13131–13141
3. Kaiwen C, Kumar A, Xavier N, Panda SK (2016) An intelligent home appliance control-based on WSN for smart buildings. In: 2016 IEEE international conference on sustainable energy technologies (ICSET). IEEE, pp 282–287
4. Rault T, Bouabdallah A, Challal Y (2014) Energy efficiency in wireless sensor networks: a top-down survey. *Comput Netw* 67:104–122
5. Maraiya K, Kant K, Gupta N (2011) Wireless sensor network: a review on data aggregation. *Int J Sci Eng Res* 2:1–6
6. Harb H, Makhoul A, Laiymani D, Jaber A, Tawil R (2014) K-means based clustering approach for data aggregation in periodic sensor networks. In: 2014 IEEE 10th international conference on wireless and mobile computing, networking and communications (WiMob). IEEE, pp 434–441
7. Harb H, Makhoul A, Couturier R (2015) An enhanced K-means and ANOVA-based clustering approach for similarity aggregation in underwater wireless sensor networks. *IEEE Sens J* 15:5483–5493
8. Idrees AK, Al-Yaseen WL, Taam MA, Zahwe O (2018) Distributed data aggregation based modified K-means technique for energy conservation in periodic wireless sensor networks. In: 2018 IEEE Middle East and North Africa communications conference (MENACOMM). IEEE, pp 1–6
9. Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, pp 1027–1035

10. Lago P, Lang F, Roncancio C, Jiménez-Guarín C, Mateescu R, Bonnefond N (2017) The ContextAct@ A4H real-life dataset of daily-living activities. In: International and interdisciplinary conference on modeling and using context. Springer, Cham, pp 175–188

Indoor Positioning System Using Pyramidal Beacon in Mobile Augmented Reality



Hyeon-woo An and Nammee Moon

Abstract This paper presents an accurate and fast indoor positioning system using beacons and a server that manages pyramidal beacons in a mobile augmented reality environment. The proposed system consists of a beacon shaped for easy recognition in the image, a detector for recognizing such a pyramidal beacon and extracting 6 degrees of freedom relative to the beacon, and a management server that manages the ID and location information of each pyramidal beacon. Because of the characteristics of the proposed method, the posture estimation including the position is also performed in the positioning. Therefore, it is possible to implement the augmented reality with a high degree of immersion and to manage a plurality of pyramidal beacons, thereby positioning them in a wide space like a large shopping mall. To manage a large number of beacons, an LED for identifying an ID exists in the center of the beacon, and the management server maps the beacon to the coordinates of the room. The positioning is performed by recognizing the pyramidal beacon through the model learned by deep learning in the mobile device, extracting 6 degrees of freedom relative to the pyramidal beacon recognized through the image-based technology, and receiving the indoor position through communication with the server. The proposed system can be applied to a user guidance service based on mobile augmented reality in a complex and wide indoor environment.

Keywords In-door positioning · Augmented reality · Mobile positioning

1 Introduction

It is an important goal to estimate the position of an object in technologies that combine the real and the virtual, such as digital twins, virtual reality, and augmented reality. Recent positioning methods using various technologies have been studied, such as image based, sensor based, finger printing, and cell ID method which use visible light, positioning with high accuracy is possible by covering and utilizing

H. An · N. Moon (✉)

Department of Computer Engineering, Hoseo University, Asan, Republic of Korea
e-mail: nammee.moon@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_3

various technologies according to the situational context. However, in the case of mobile devices mainly used for augmented reality, there is no navigation system with an accuracy high enough to utilize dead reckoning, and there are not enough sensors to accommodate various methods unless they are specially manufactured for positioning. Even if the positioning method is possible in the mobile environment, such as with GPS or WiFi Fingerprinting, it is difficult to implement augmented reality with high immersion owing to inaccurate results.

In this paper, we present a system that can overcome the limitations of mobile devices for positioning in augmented reality and enable highly accurate positioning. The system detects a fixed form of pyramidal beacon using an artificial neural network in a mobile augmented reality environment and calculates the relative position of the user by utilizing the image of the detected beacon. Thus, more accurate and immersive augmented reality-based information can be provided; moreover, the marker-based stable information can be provided.

2 Related Works

2.1 *Multiple Object Detection*

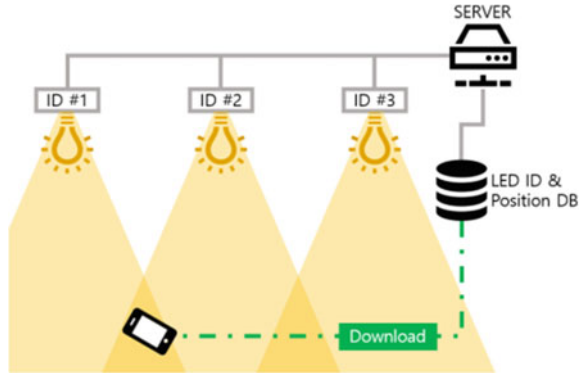
The method of detecting objects in the image is largely divided into an algorithm-based method using the image characteristics and a method using a learned neural network. The algorithm-based method using image qualities has a limit that cannot be individually reflected in the situations where there is a possibility of false positives, and there is a disadvantage that the detection method may need to be modified as a whole according to the change of the target object. Contrarily, in the case of the artificial neural network-based method, it is possible to reduce false positives according to the quantity and quality of the learning data, and to efficiently detect bad conditions.

There are Faster R-CNN, YOLO (You Only Look Once), R-FCN (Region-based Fully Convolutional Network), SSD (Single Shot Detector), etc., in the multi-object detection method based on the artificial neural network, and each detection method generally has inherent inaccuracy and is inversely proportional to the fps (frames per second) [1–3]. In this study, we decided to use YOLO v3, which has a faster detection speed than other methods, because the pyramidal beacon should be detected close to real time in mobile environments [4].

2.2 *Mobile-Based Indoor Positioning*

Several well-known indoor positioning methods that can be used in mobile devices include WiFi Fingerprinting, BLE beacon, and the camera-based methods.

Fig. 1 Light based positioning. This method emits light at independent frequencies in each led bulb, and the imaging device uses the frequency of the captured led bulb as an ID to identify the cell where it is currently located



The WiFi Fingerprinting scheme is a method of constructing a table (a radio map) mapping MAC addresses and absolute positions to three or more APs and measuring the received signal strength with each AP of the device through a trilateration [5]. In this case, the intensity of the signal changes depending on various factors such as the device state, temperature, humidity, and obstacles. Consequently, detailed positioning is difficult unless it is a special situation.

The positioning method using the BLE beacon is based on the intensity of the signal, similar to the WiFi Fingerprinting method [6]. In addition, it is possible to perform trilateration by the indoor positioning method. However, it is difficult to precisely locate it because it is possible to measure the approximate position by the influence of various interferences.

Finally, with the improvement of the camera performance and processing speed of the mobile device, the method using the image can be positioned with a relatively higher accuracy than the previous two methods. For example, there is a technique of locating the front camera data of a mobile device by adjusting the blinking of each LED so that it can not be perceived by a person [7]. This is applicable to immersive augmented reality implementation because the error is in units of cm. However, the camera device must be shooting seamlessly towards the LED Bulb and must capture two or more LED Bulbs in the image for more accurate position measurements. Figure 1 illustrates the VLC based positioning method described above.

3 Pyramidal Beacon-Based Positioning System

The system can be largely divided into the learning part, the augmented reality computing part, and the pyramidal beacon management server part. The learning part learns the beacon image for area detection by using YOLO, and the learned model is used for area detection in the augmented reality operation part. The augmented reality operation part aims to output augmented information based on the camera input data in a mobile environment. The pyramidal beacon management server manages the

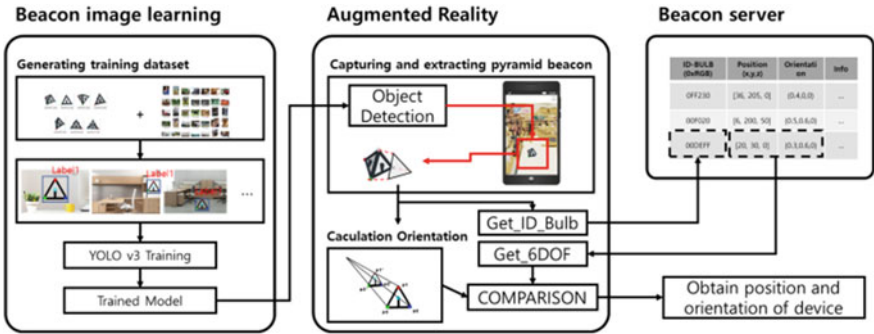


Fig. 2 System overview

absolute coordinates of each beacon to estimate the relative distance between the beacon and the mobile device, while mapping the beacon information with the ID bulb for providing information. The Fig. 2 shows the system overview and outlines the overall process up to the positioning.

3.1 Pyramidal Beacon

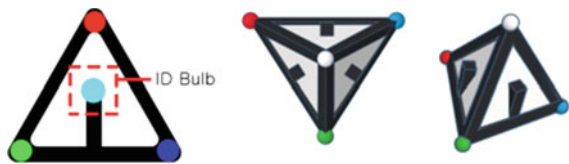
The structure of the pyramidal beacon is shown in the Fig. 3 below. To more easily detect geometric features in the image, LEDs of different colors are present at each of the three vertices, and the ID bulb for identifying the beacon is included in the center. The ID bulb can be split and assigned to a mobile device recognizable unit in the color spectrum.

Estimating the transformed orientations from the captured beacons is performed using homography. Homography is the most general model that can explain the transformation relation of planar objects, and it is suitable for measuring the rotational state of the pyramidal beacon [8].

In this study, YOLO v3 was adopted as a method of detecting on pyramidal beacons because it is most important to reflect on the mobile device in real time. The learned model operates in a mobile application, and it is used to detect the area of a pyramidal beacon.

The data set for training is generated by combining the images of the beacons (as foreground images) and the VOC data sets (as background images). The training is

Fig. 3 The pyramidal beacon model drawing



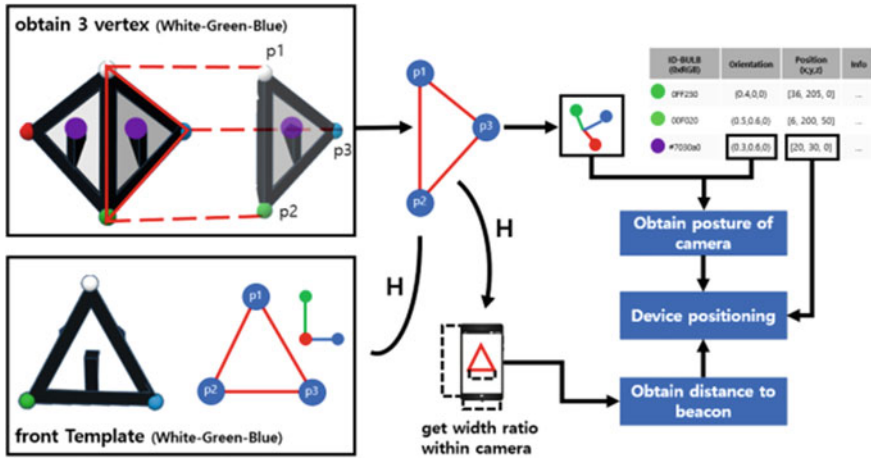


Fig. 4 The entire process of positioning

carried out through about 10,000 generated images; consequently, a model is created with an average loss value of 0.06, enabling the detection of beacon areas in the image.

3.2 Pyramidal Beacon Management Server

The pyramidal beacon management server maps each beacon location and ID bulb value, and the information of each pyramidal beacon can be added.

For relative positioning with installed pyramidal beacons, the information about orientation (roll, pitch, and yaw) is needed, as well as the information about position. Therefore, six pieces of state information describing the location are stored and matched with the ID bulb.

When the ID of the pyramidal beacon is detected by the mobile device, the bulb color is transmitted to the server and; the server refers to the ID bulb value in the managed radio map and returns the corresponding beacon information. The ID bulb color values may vary depending on the performance of the illumination or the camera. To compensate for the ID bulb color values, the color values of the three vertices are received and compared with the predetermined color values, thereby helping to identify the ID Bulb color more clearly.

3.3 Positioning

Positioning of mobile devices is shown in the Fig. 4 and consists of the below processes: First, pyramidal beacon detection is performed based on the camera input data of the mobile device. As described above, a YOLO v3 model that has been learned in advance is utilized for detection. If the detection proceeds normally, then the orientation is obtained through the homography operation on the detected pyramidal beacon. At this time, one surface having three vertices (p_1, p_2, p_3) is extracted from the beacon to obtain a changed rotation state of the beacon through homography and subsequently compared to the front state image that was previously stored for the surface. In the following process, the captured beacon face is converted to the frontal state using the H matrix obtained through homography, and the relative distance to the beacon is obtained by comparing the zoom value and the aspect ratio of the photographing device.

The management server then uses the ID-Bulb color value of the detected pyramidal beacon to obtain the 6-DOF of the beacon. This 6-DOF consists of the absolute coordinates and orientation of the pyramidal beacon installed in the indoor space. Contrasting the obtained 6-DOF with the orientation of the captured pyramidal beacon can create a vector describing the orientation relationship of the mobile devices in the indoor space. Then, the captured region is transformed into the frontal state using the homography property of the pyramidal beacon, and the distance between the pyramidal beacon and the mobile device is calculated by obtaining the scale conversion. At this time, because the conversion relation of the scale can not be explained through simple image-based contrast, the camera profile information of the mobile device is utilized.

4 Conclusion

In this paper, we proposed a system for real-time indoor positioning in a mobile augmented reality environment. For this purpose, the YOLO v3 was used to capture region of beacon in real time, and homography was used to explain the transformation relation of the beacon in image. The proposed system is similar to the marker based on the ID bulb included in the pyramidal beacon, but it can be managed more flexibly, and will help to realize immersive augmented reality. The system is designed to only be capable of positioning when a beacon is captured within the camera of the mobile device. Therefore, in the case of functions that require continuous positioning such as route finding, it is considered necessary to develop more advanced techniques such as auxiliary positioning of a space where no beacon is captured through dead reckoning.

Acknowledgements This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. NRF-2017R1A2B4008886).

References

1. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*, pp 91–99 (2015)
2. Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. In: *Advances in neural information processing systems*, pp 379–387 (2016)
3. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: *European conference on computer vision*. Springer, Cham, pp 21–37
4. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
5. Haeberlen A, Flannery E, Ladd AM, Rudys A, Wallach DS, Kavraki LE (2004) Practical robust localization over large-scale 802.11 wireless networks. In: *Proceedings of the 10th annual international conference on Mobile computing and networking*. ACM, pp 70–84
6. Faragher R, Harle R (2015) Location fingerprinting with Bluetooth low energy beacons. *IEEE J Sel Areas Commun* 33:2418–2428
7. Kim HS, Kim DR, Yang SH, Son YH, Han SK (2012) An indoor visible light communication positioning system using a RF carrier allocation technique. *J Lightwave Technol* 31:134–144
8. Sukthankar R, Stockton RG, Mullin MD (2001) Smarter presentations: exploiting homography in camera-projector systems. In: *Proceedings eighth IEEE international conference on computer vision*. ICCV 2001, vol 1. IEEE, pp 247–253

Design and Implementation of Real-Time Vehicle Recognition and Detection System Based on YOLO



Hyeonmoo Jeon, Gilwoo Lee, Byeongcheol Jeong, Jae Sung Choi, Jeong-Dong Kim, and Bongjae Kim

Abstract Researches are continuously carried out to provide various services based on the recognition of vehicles and their recognition results. For example, related technologies can be used in various intermitted cameras. In this paper, we proposed a real-time vehicle recognition and detection system based on YOLO. The proposed method is designed and implemented to recognize vehicles using CNN based on YOLO. The proposed method has the advantage of recognizing the vehicle number and the type of vehicle at the same time, and it can be applied to applications such as the prevention of various crimes using vehicles.

Keywords Vehicle recognition · License plate recognition · Deep learning · YOLO

1 Introduction

A variety of services can be provided based on the recognition of the license plate of the vehicle. Typical applications include various crackdown services such as a crackdown on illegal parking regulations. In such applications, recognizing vehicle

H. Jeon · G. Lee · B. Jeong · J. S. Choi · J.-D. Kim · B. Kim (✉)
Division of Computer Science and Engineering, Sun Moon University, 70, Sunmoon-ro 221
Beon-gil, Tangjeong-myeon Chungcheongnam-do, Asan-si 31460, South Korea
e-mail: bjkim@sunmoon.ac.kr

H. Jeon
e-mail: sever5619@gmail.com

G. Lee
e-mail: dlrfdnwd12@gmail.com

B. Jeong
e-mail: jbilly8349@gmail.com

J. S. Choi
e-mail: jschoi@sunmoon.ac.kr

J.-D. Kim
e-mail: kjd4u@sunmoon.ac.kr

license plates is a very important problem because they operate their services based on license plate recognition results. However, license plate recognition alone is not enough because there may be vehicles whose license plates are stolen or changed illegally.

In this paper, we propose a real-time vehicle recognition and detection system to solve this problem. The proposed vehicle recognition and detection system can recognize the license plate and the type of vehicle at the same time. The proposed vehicle recognition and detection system is designed and implemented based on the YOLO (You Look Only Once) [1]. The prototype of our vehicle recognition and detection system provides a Web-based management tool to manage and monitor the detection and recognition results. The average recognition accuracy of the vehicle type was about 87%, and the recognition accuracy of the license plate was about 90%. We think that if we do more study to improve the performance further, the proposed system can be effectively used for preventing various crimes using vehicles.

The rest of this paper is organized as follows. In Sect. 2, some related works will be described. In Sect. 3, we will explain the structure and service flow of the proposed real-time vehicle recognition and detection system. In Sect. 4, we will show the implementation results and its performance in terms of recognition accuracy of the license plate and vehicle type. Finally, we conclude this paper with future works in Sect. 5.

2 Related Works

Typical vehicle identification systems based on image processing techniques consist of three stages: license plate detection, character segmentation, and character recognition. The process of license plate detection is the most important part of the three stages. This is because other processes proceed after the detection of a license plate. Therefore, if the license plate is not recognized correctly, it is difficult to recognize the number constituting the license plate. In order to alleviate this problem, studies related to license plate detection and recognition based on deep learning are also being carried out [2, 3]. Apart from this, vehicle recognition and detection systems for traffic surveillance systems are also being studied [4, 5].

3 Proposed Real-Time Vehicle Recognition and Detection System

Figure 1 shows a service flow of the proposed vehicle recognition and detection system. The proposed system has two major components. The First one is YOLO based real-time vehicle detection recognition component. The last one is the Web-based management tool.

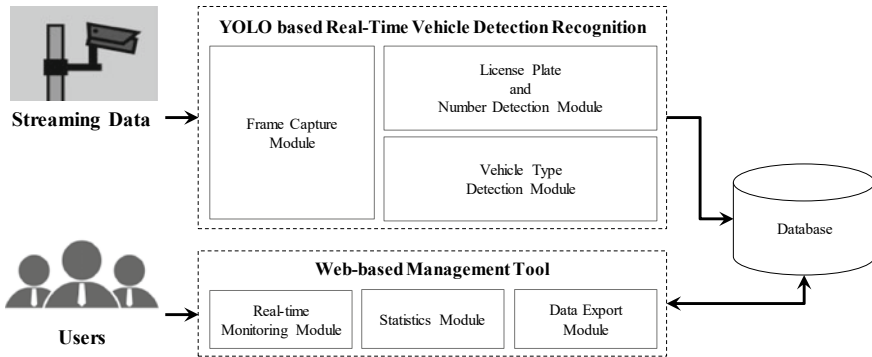


Fig. 1 A service flow of the proposed vehicle recognition and detection system

The service flows are as follows. As shown in Fig. 1, streaming video data is received. The received streaming data is divided into frames by the Frame Capture Module (FCM). Then, the License Plate and Number Detection Module (LPNDM) analyzes each frame data and recognize the car number by using YOLO. Similarly, the Vehicle Type Detection Module (VTDM) recognizes the type of vehicle. Recognition results are stored in the database in real-time.

Vehicle recognition and detection functionalities are implemented based on YOLO v3. In our system, we modified the size of the input image when using YOLO. The image sizes for vehicle identification and number recognition are set differently. 832×832 pixel image was used for vehicle type identification to train and test the model. Similarly, 608×608 pixel image was used for recognizing the vehicle number to train and test the model. This is a parameter that shows the best performance from the experimental results.

The stored recognition results are visualized in real-time by the Web-based management tool and can be confirmed by the user or administrator. The Web-based management tool supports three major functionalities: real-time monitoring, statistics, and data export functionalities.

4 Implementation Results and Performance Evaluations

Figure 2 shows an example of the proposed real-time vehicle recognition and detection system. Figure 3 shows an example of the detection and recognition results of the proposed system. As shown in Figs. 2 and 3, the user and administrator can manage and monitor detection and recognition results via the Web-based management tool. As shown in Fig. 2, the location information on which the vehicle is recognized and detected is displayed on the map when the officially registered information of the vehicle is different from the recognized vehicle information.

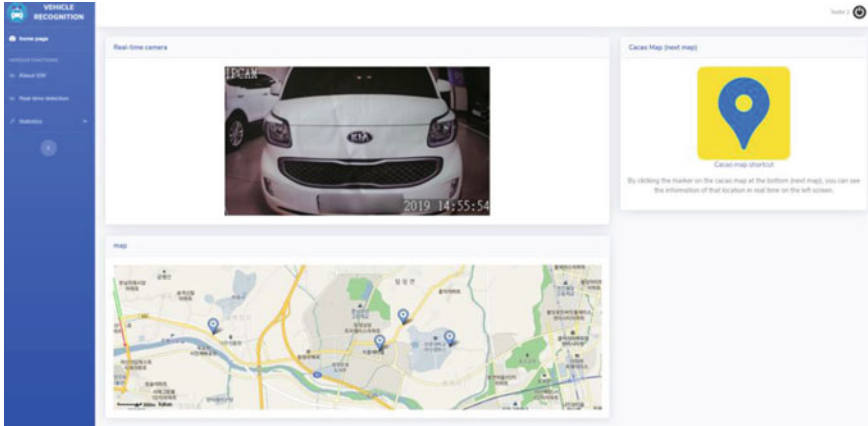


Fig. 2 An example of the proposed real-time vehicle recognition and detection system

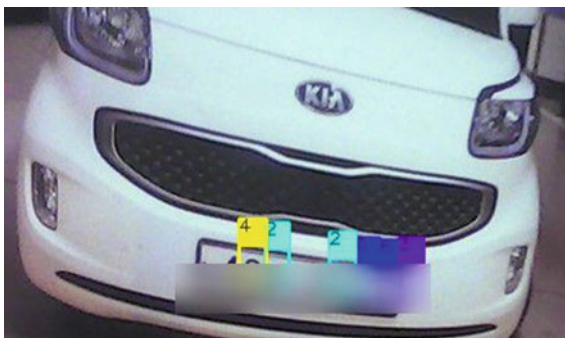
Index	Type of crime	Location	Vehicle model	the car's number	Recognized time	type	Plate
0	Counterfeit vehicle	Sun Moon University	ray	412412	2019-06-09 20:50:00	vehicle	number
1	Counterfeit vehicle	Sun Moon University	ray	921921	2019-06-09 20:50:46	vehicle	number
2	Counterfeit vehicle	Sun Moon University	ray	356356	2019-06-09 20:50:54	vehicle	number
3	Counterfeit vehicle	Sun Moon University	santafe	662662	2019-06-09 20:51:15	vehicle	number
4	Counterfeit vehicle	Sun Moon University	starex	633632	2019-06-09 20:51:26	vehicle	number
5	Counterfeit vehicle	Sun Moon University	santafe	463846	2019-06-09 21:23:14	vehicle	number
6	Old car	Sun Moon University	ray	422663	2019-06-09 21:35:09	vehicle	number

Fig. 3 An example of detection and recognition results of the proposed system

Our prototype can recognize five types of vehicles: Hyundai I40, Hyundai Morning, Hyundai Santa Fe, Hyundai, Starex, and Kia Ray. The average precision of the detection of vehicle type was about 87%.

Figure 4 shows an example of the detection and recognition result of a license plate and its number. In Fig. 4, the license plate of the vehicle was blurred because it contains personal information. The average precision of detection of license plates and its number was about 90%.

Fig. 4 An example of detection and recognition result of a license plate and its number



5 Conclusions and Future Works

In this paper, we proposed a real-time vehicle recognition and detection system based on the YOLO framework. The proposed system can detect and recognize the number of a license plate and the type of vehicle at the same time. Because of this advantage, it is possible to detect vehicles which have different license plate number from officially registered vehicle information in real-time. Based on the performance evaluation results of our prototype, the average recognition accuracy of the vehicle type was about 87%, and the recognition accuracy of the license plate was about 90%. In the future works, we will focus on increasing the recognition accuracy and the number of types of vehicles that can be recognized in the system for practical use.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1C1B5017476), supported by Institute for Information & communications

Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2018-0-01,865), and supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5045953).

References

1. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
2. Polishetty R, Roopaei M, Rad P (2016) A next-generation secure cloud-based deep learning license plate recognition for smart cities. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 286–293
3. Silva SM, Jung CR (2017, October) Real-time Brazilian license plate detection and recognition using deep convolutional neural networks. In: 2017 30th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 55–62

4. Tang Y, Zhang C, Gu R, Li P, Yang B (2017) Vehicle detection and recognition for intelligent traffic surveillance system. *Multimedia Tools Appl* 76(4):5817–5832
5. Betke M, Haritaoglu E, Davis LS (2000) Real-time multiple vehicle detection and tracking from a moving vehicle. *Mach Vis Appl* 12(2):69–83

Purchase Predictive Design Using Skeleton Model and Purchase Record



Jae-hyeon Cho and Nammee Moon

Abstract The depth camera has enabled the skeleton and joints of the human body to use skeleton data in 3D space. Behavior recognition using skeleton data is mainly based on artificial neural networks such as RNN. This study classifies behaviors observed by the consumer into four categories using skeleton model learning for purchase predictive design. Skeleton model learning collects 25 skeleton joints using several Kinect v2s in unattended stores where four racks of items can be purchased. Torso, left arm, right arm, left leg, and right leg to five body joints are performed by BRNN, and as the layer becomes deeper, each part is then joined to the body. Finally, the 25 joints are grouped together and BRNN-LSTM is performed to solve the vanishing gradient problem (Jun et al. in *J Korea Multimedia Soc* 21:369–381, 2018, [1]). Supervised learning involves four behaviors used as input and the purchase record status as output. A GRU is employed to reduce computational complexity while maintaining the benefits of LSTM.

Keywords BRNN · Purchase prediction · Skeleton model

1 Introduction

Behavior awareness can be used in various fields such as the detection of dangerous behavior, robot vision, and game control, and is considered an important field in computer vision research today. The Kinect sensor is a device that provides the ability to track the human skeleton based on acquired depth map information, thereby making it possible to construct a low-cost non-contact motion capture system.

As the behavior of a person is determined by the progress of movement according to the passage of time, the time series problem must be considered to improve the

J. Cho · N. Moon (✉)

Division of Computer and Information Engineering, Hoseo University, Asan, South Korea

e-mail: nammee.moon@gmail.com

J. Cho

e-mail: a01032629198@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_5

accuracy of behavior recognition. Recurrent Neural Network (RNN) is a machine learning technique for learning and solving problems affecting data in the previous sequence in the time series. Recently, HBRNN has been proposed, where joints in skeleton data are divided into five body parts including torso, left arm, right arm, left leg, and right leg to generate RNNs and hierarchically combine those corresponding to body parts in close proximity to each other [2, 3].

In early 2018, Amazon opened the San Francisco grocery store Amazon Go. Amazon Go is designed for people who prefer not to stand at the checkout counter. Just Walk Out Technology offers a more convenient service by tracking purchases with sophisticated technology, where consumers leave the store with purchases automatically charged to their account. This technique is similar to that used in autonomous vehicles. The foundation is backed by cutting-edge computer technology such as computer vision, sensor fusion, and deep learning. Multiple cameras are recognized and tracked individually from the moment the customer enters the shop. Microphones, pressure sensors, and weight sensors installed at each stand record every move of the purchaser. Payment is made by actively utilizing the propensity of the purchaser who has been confirmed and past purchase data [4].

In this study, we used Kinect with a built-in RGB camera and a depth camera to classify behaviors using skeleton model learning. Using this behavior and the consumers past purchase records, purchase predictive learning is performed to determine whether or not the consumer will purchase the product.

2 Related Work

In this study, it is important to maximize the collection rate of Skeleton Joints and the accuracy of learning between Skeleton Joints. In this session, you will find out the BRNN used for learning from Kinect and Skeleton Model Learning, who are responsible for Skeleton Tracking.

2.1 *Kinect*

Kinect is a peripheral device that connects with the XBOX 360, a gaming console and entertainment system that detects input via gestures from the human body instead of using the traditional gaming controller. Kinect recognizes user actions through the sensor and recognizes voice using the microphone module. Kinect sensors are low-cost depth cameras that provide real-time depth information as well as RGB images and joint tracking information. Data provided by the Kinect sensor eases the effort involved in human body part detection and pose estimation necessary for gesture recognition, making it easier to develop human-computer interaction applications.

2.2 BRNN

BRNN is a type of deep learning algorithm based on the existing RNN. The BRNN algorithm is composed of an input layer, a forward layer, a backward layer, and an output layer. The forward and backward layers act as the hidden layer of the RNN algorithm [5].

3 Method

Figure 1 presents a flow chart of our system design. Kinect collects the consumer's behavior in front of the rack to obtain a skeleton model. The consumer record information is used to obtain the behavior and purchase status in the experimental environment. The skeleton model learns by classifying behaviors into four specific categories using BRNN. In the course of supervised learning, this result is learned so that purchaser record information is purchased or not.

3.1 Data Set

Shoot people with Kinect to collect data. The skeleton tracker measures data such as the height, angle and distance and this information is used to recognize a person, and to select the location with the lowest error value. The pygame module is used to receive both the RGB and depth maps at the same

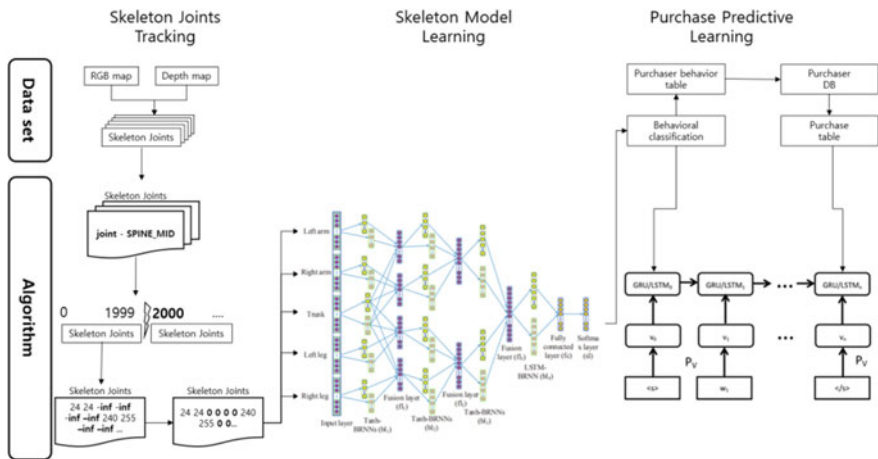


Fig. 1 System design flow chart

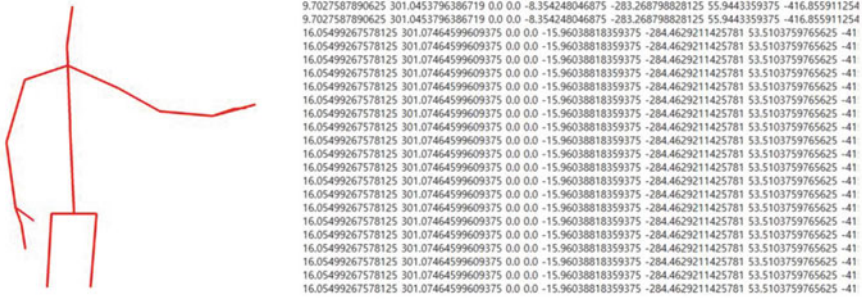


Fig. 2 Skeleton model collection

time. A total of 25 joints are collected by the Skeleton Tracker, these include: SpineBase, SpineMid, Neck, Head, ShoulderLeft, ElbowLeft, WristLeft, HandLeft, ShoulderRight, ElbowRight, WristRight HandRight, HipLeft, KneeLeft, AnkleLeft, FootLeft, HipRight, KneeRight, AnkleSight, FootRoot, HandTipLeft, ThumbLeft, HandTipRight, ThumbRight. When saving, 50 pieces of data are accumulated per frame by storing the x and y coordinates of each part in order in each frame. The data is stored by converting the model on the left into coordinates on the right, as shown in Fig. 2. Each action is represented by a single action to make learning easier.

The 25 stored skeleton joints go through three preprocessing steps before being used as input data for learning. The first step is to convert absolute coordinates to relative coordinates. Skeleton joints subtract SpineMid values so that the values do not change wherever a person is located in the image. SpineMid represents the center of the body and the same relative coordinates appear if a person behaves in the same way even if they are located at different coordinates. The second process is to cut the frame to a certain length, and this is achieved by storing data from the first frame to the 2000th frame. When shooting, certain joints are not collected due to occlusion, overlapping, or poor recognition, resulting in an error value, so the third step is to correct the error value.

3.2 Skeleton Model Learning

Preprocessed data are used as input data for skeleton model learning. In 2000 frames, 25 x, y coordinates are used as data of the form [2000 * n, 50]. Four folded motions are used as output data as shown in Fig. 3. These are pick-up motion, drop motion, arm-twisted motion and motion that brings one’s face to an object. Motion is expressed as data of [2000 * n, 4] form. Of the data collected, 80% is used for training and 20% is used for test data.

To teach the skeleton model, we use HRNN, which consists of three mixed layers and 1 BRNN-LSTM layer. A mixed layer consists of learning incoming data using

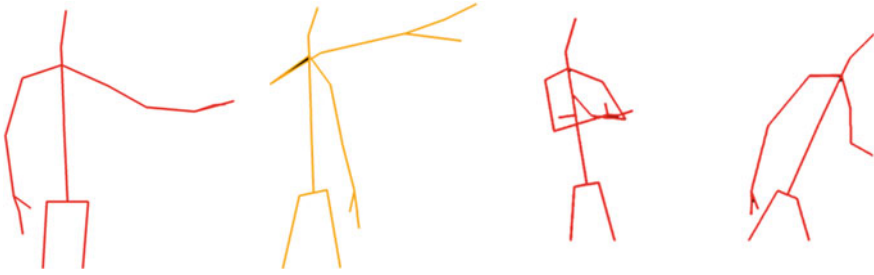


Fig. 3 Four skeleton behavioral classifications

BRNN and combining these results. The coordinates obtained during the data collection process are grouped into the torso, left arm, right arm, left leg, and right leg. These data are used as the input value and BRNN is performed for each. BRNN is then performed in combination with the torso and the remaining joints, and finally BRNN-LSTM is performed in its entirety. The results are then classified into four behaviors using Softmax.

The first mixed layer consists of five BRNNs. To model adjacent body parts, the torso group is combined with the other four to obtain four new groups from the convergence layer. The four groups created are composed of four BRNNs, as is the first blending layer. The left arm and right arm groups are combined to model the upper body, while the left foot and right foot groups are combined to model the lower body. Finally, the two groups obtained from the previous layer are composed of two BRNNs and reassembled to represent the entire body. Time series representations throughout the body consist of BRNNs, which are different from previous layers. Once the final feature of the skeleton model is obtained, the fully connected layer and the Softmax layer are performed to classify the behavior appropriately [6].

3.3 Purchase Predictive Learning

Purchase prediction executes the supervised learning process by setting the consumers purchase history as a result value with GRU used for purchase forecasting. Recent comparisons with LSTM, which is widely used as a circular neural network, have shown higher accuracy as the amount of data increases. The result of the behavior classification learning is used as the input value of the GRU and the result is 1 if the purchaser purchased it and 0 if the purchaser did not purchase it [7].

4 Conclusion

Although Amazon Go has an innovative unmanned store model, the Korean retail industry is focused on reducing costs through unmanned payment technologies. Related technology development is also focused on building payment systems and security technologies rather than data collection. In order for future similar stores to emerge in Korea, it is necessary to analyze past purchasing patterns of individual consumers and collect high quality data.

This study proposes a model that predicts consumer purchases using Kinect and existing purchase records. We have shown that it is possible to obtain good quality data at relatively low cost in predicting the purchase pattern of consumers. Future work will establish how this research can support unattended stores emerging in Korea that lack data collection technology.

Acknowledgements This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1A2B4008886).

References

1. Jun J, Hwang S, Yoon Y (2018) A verification about the formation process of filter bubble with personalization algorithm. *J Korea Multimedia Soc* 21(3):369–381, 13
2. Sang U, Park K, Lee Y-K (2017) Human activity recognition based on RNN with using correlations in skeleton data. *The Korean Institute of Information Scientists and Engineers*, pp 755–757
3. Kim M-K, Cha E-Y (2018) Using skeleton vector information and RNN learning behavior recognition algorithm. *J Broadcast Eng* 23(5):598–605
4. Polacco A, Backes K (2018) The Amazon go concept: implications, applications, and sustainability. *J Bus Manage* 24(1):79–92
5. Kim A-R, Rhee S-Y (2018) Recognition of natural hand gestures using bidirectional long short-term memory model. *Int J Fuzzy Logic Intell Syst* 18(4):326–332
6. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp 1110–1118
7. Graves A, Mohamed A-R, Hinton G (2013) Speech recognition with deep recurrent neural network. In: *2013 IEEE international conference on acoustics, speech and signal processing*

Intelligent Digital Signage Using Deep Learning Based Recommendation System in Edge Environment



Kihoon Lee and Namme Moon

Abstract This paper proposes a highly intelligent digital signage system based on the IoT (Internet of Things) edge equipped with a learned advertisement recommendation model. The proposed system consists of a server and an edge. The server manages the data, learns the advertisement recommendation model, and the edge determines the advertisement to be promoted in real time using the learned advertisement recommendation model. The ad recommendation model consists of a selection of products and a prediction of their purchasing probabilities. In the screening phase, the user-based information and product metadata vectored into DNN are entered to derive the product that is worth purchasing. We use a soft max function to predict the purchase probability of selected goods. Finally, the most suitable advertisement is selected by using the predicted purchase probability of the community. The proposed system does not communicate with the server. Therefore, it decides the advertisement with the learned model on the edge. This also applies to digital signage that requires immediate response to many users.

Keywords Edge computing · Digital signage · Deep learning · Recommended system

1 Introduction

Digital signage provides various forms of services to various displays or billboards using digital technology for advertising in public and commercial spaces. Digital signage works by serving various media contents to various indoor and outdoor displays. With the continued development of IoT technology and ICT, digital signage has evolved with various technologies in the fourth industry [1]. In addition, in outdoor advertising, which simply leads to information transmission, it is possible to interact with consumers by using given real-time information. The value has been increasing for a service that is friendlier and more interactive with consumers [2].

K. Lee · N. Moon (✉)

Department of Computer Engineering, Hoseo University, Asan, Korea
e-mail: nammeemoon@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_6

In ISE 2018, Samsung introduced “Target Ad Signage Solutions,” which can be used in advertising. Using the display along with the camera, basic information such as the sex and age of individuals within the advertisement range are identified using artificial intelligence to display the customized advertisement. The display conveys one-sided information of a person and also provides a customized view so that the user can see the related content that the user wants to view. The core of such intelligent digital signage is artificial intelligence technology. An area of AI, consumer behavior prediction technology based on machine learning, is becoming more sophisticated, and large amount of data produced daily is used as a basis for predicting consumer behavior. The key of extracting meaningful information from the data is the artificial neural network. The accuracy of the learning result depends on the type of layer constituting the artificial neural network and the learning data [3]. It has been shown that the accumulation of massive amounts of data and the development of rapid semiconductor technology, which is called the Big Data Age, can considerably improve the accuracy of artificial neural networks [4, 5].

Previously, artificial intelligence was commonly used to transfer all data to the data center or the cloud [6]. However, recently, edge computing is moving to change artificial intelligence to the edge [6]. Edge computing does not process data centrally but processes data directly at the edge where the data is generated. This allows the user to analyze and utilize the data in real time by reducing the traveling time of the data to the cloud [7]. The most active field of edge computing is autonomous vehicles. The autonomous vehicle produces about 4 TB of data per hour from various sensors. Autonomous vehicles are required to analyze and process data continuously generated during driving in real time, which is directly related to passenger safety [8]. For this, we analyze and judge data in real time using artificial intelligence in vehicles with sensors, that is, the edge, and deal with the various situations. Edge computing has the advantage of minimizing the delay time by quickly analyzing and determining data in the module where the data is directly generated [9, 10].

In this paper, we propose an intelligent digital signage system that selects the most suitable advertisement for the user’s community in the current situation through an artificial intelligence model in the edge environment. We estimated the purchasing probability of each user by using the soft max multiple classification function. The score for each product is derived using the derived purchase probability. The user’s past purchasing history, current weather, and user-customized product list are derived based on the score. The final advertisement list is selected by considering the product list of all users within the advertisement range. We propose an intelligent digital signage system in an unmanned shop by applying this edge-based system to the digital signage that must interact with the user in real time.

2 System Overview

This system proposes an intelligent digital signage system using edge computing in the IoT environment. The proposed intelligent digital signage system is shown in Fig. 1 is divided into a server and an edge.

The server has a Product DB that stores information on items displayed such as price, product type, and product location. Moreover, it has a data center that stores

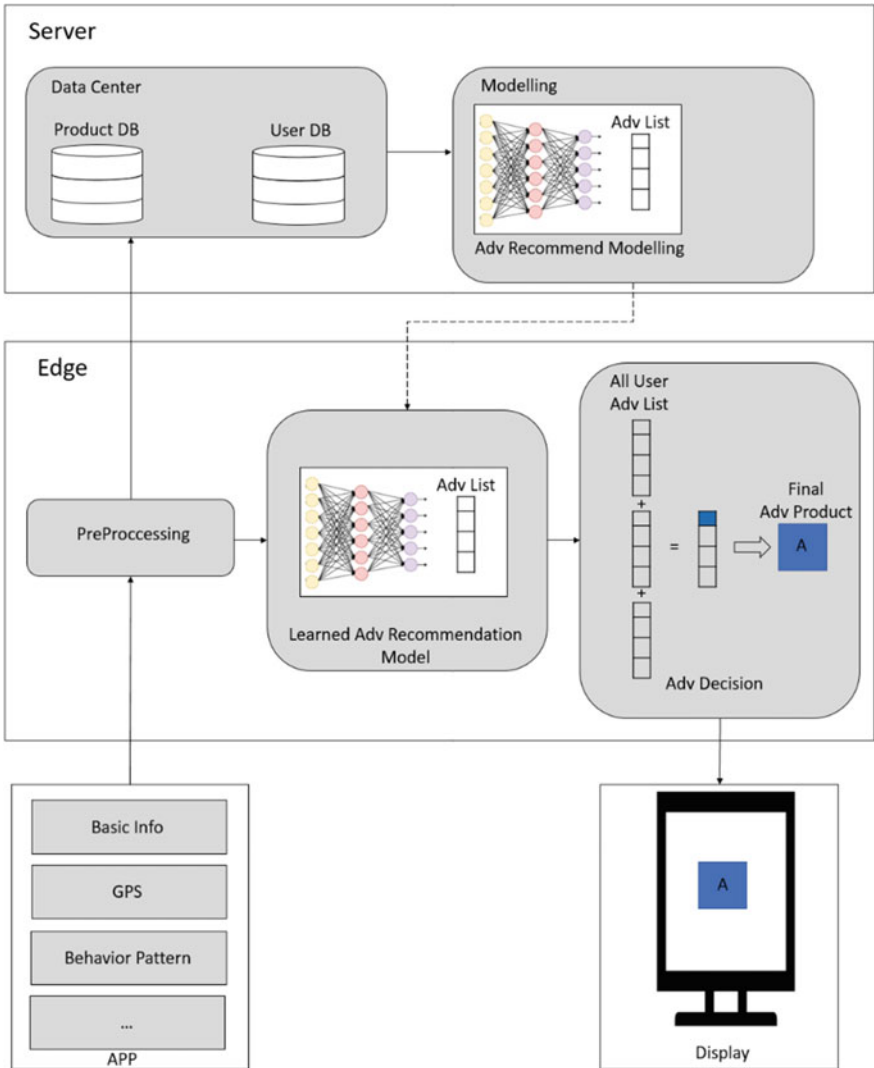


Fig. 1 System overview

the user’s purchase history and user information. In addition, we use the data in the data center to learn product metadata and users’ purchasing history to develop an ad recommendation model.

In the edge, raw data obtained from the application and various sensors in real time is used as the input values for the server and recommended model in the preprocessing process. In addition, by using the user information and weather data as the input values, the advertisement recommendation model derives the recommendation list for each user. The final advertisement item is selected considering the purchase probability list of all users in the advertisement range.

Finally, the advertisement for the recommended article is used to reproduce the advertisement to the user through the in-store display. After playback, it grasps the user’s behavior pattern, such as purchases, and uses it to strengthen the model.

2.1 Product Candidate Generation

The ad recommendation model consists of a two-level in-depth neural network model like Fig. 2. The reasons for making the recommendation are as follows. First, limited information is used to first narrow the data that must be analyzed by the network. This is done to use more information on the second network within the narrowed data range to accurately recommend a product that the user wants to view. The ad recommendation model consists of a product candidate generation model stage that predicts the goods to be purchased and a DNN-based recommendation model that predicts the purchase probability of the candidate’s products. Second, the product candidate group generation model embeds the basic information along with the product information of the customer and inputs it into the DNN. By using this information, the model extracts the product to be purchased by the customer among other products. In the Product Candidate Generation phase, we use the Softmax function to predict the purchase probability of a product for each user.

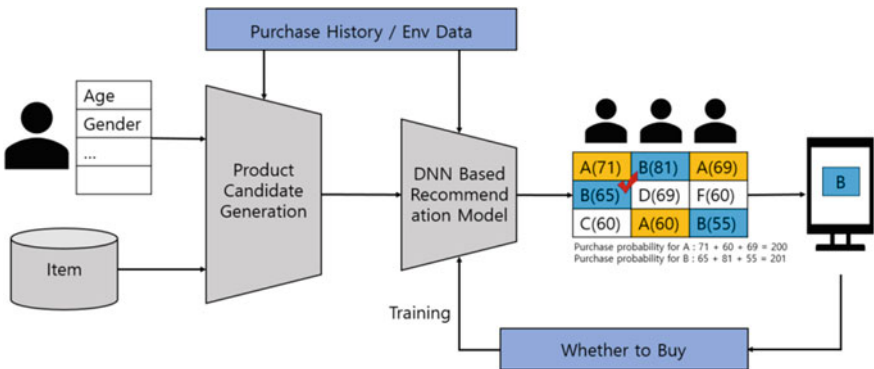


Fig. 2 Advertisement recommendation model

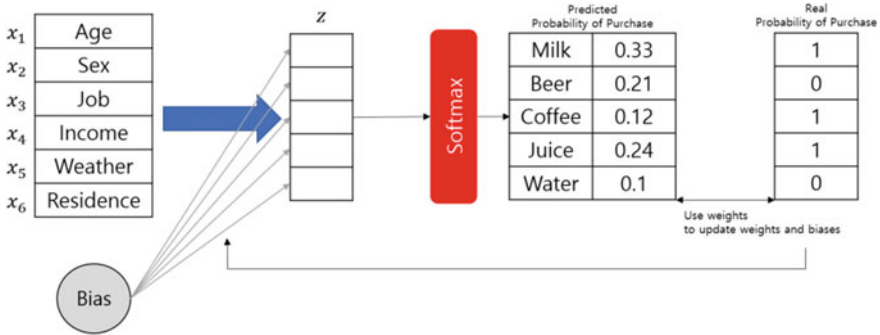


Fig. 3 Softmax function

Figure 3 shows the process of vectorizing various data such as gender, age, occupation, and income of the user through the embedding process. It also plays a role of updating weights and biases until the minimum error value is obtained by using the error between the predicted purchase probability and the real purchase probability.

And to forecast the user’s purchase probability of the selected recommendation item as shown in the equation below.

$$P_i = \frac{e^{z_j}}{\sum_{j=1}^k e^{z_j}} \text{ for } i = 1, 2, \dots k. \tag{1}$$

In the k-dimensional vector, z_j is the i-th element, and P_i is the purchase probability for the ith product.

Figure 4 is a product preference score matrix for all users within the advertising scope. Based on the above user-to-product purchase probability matrix, the product to be advertised is selected by choosing the product with the highest preference score when it is advertised to people within the signage advertisement range.

3 Conclusion

This paper proposes a highly intelligent digital signage system based on the IOT edge equipped with the learned advertisement recommendation model. The proposed system consists of a server and an edge. The server managed the data, learned the advertisement recommendation model, and Edge could determine the advertisement to be advertised in real time using the proposed model. The ad recommendation model consists of two steps: selecting products and predicting the purchase probability. In the selection step, the product that can be purchased is derived by inputting vectorized user basic information and product metadata in the DNN. Furthermore, the soft max function was used to predict the purchase probability of the selected products. Finally, the most suitable advertisement was selected using the predicted purchasing

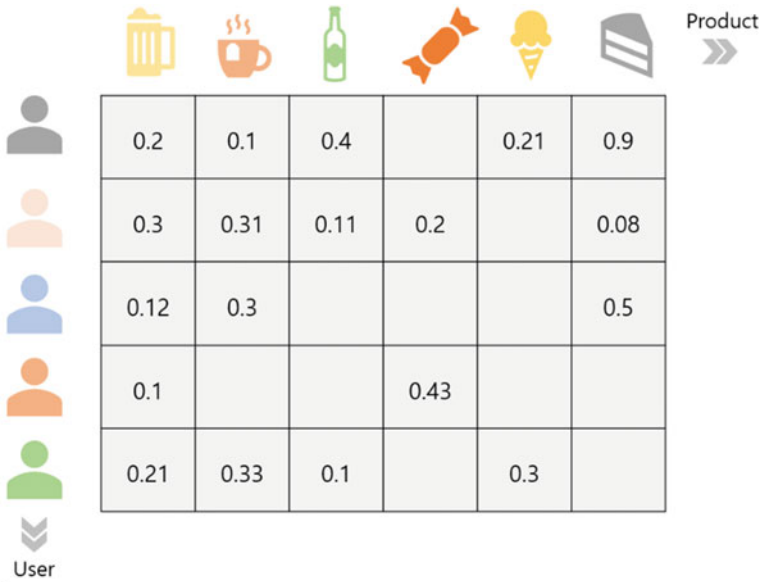


Fig. 4 Preference matrix

probability of the community. The proposed system does not communicate with the server and decides what to advertise by using the learned model on the edge. This is suitable for digital signage requiring immediate response to many users. In this study, we applied the intelligent digital signage to the uninhabited store to measure the purchase frequency of the advertisement and showed the validity of this system. In future studies, we will strengthen the model by studying the method of learning the model used and judging whether the user will purchase the product.

Acknowledgements This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1A2B4008886)

References

1. Kim EY, Sun HO (2017) Consumer emotional experience and approach/avoidance behavior in the store environment with digital signage. *J Korean Soc Cloth Textile* 41(2):266–280
2. Choi SJ, Jo YH, Sohn I (2019) Intelligent digital signage system implementation based on emotion recognition algorithm. *J Inst Electron Inf Eng* 56(3):63–72
3. Hong JP, Kim EJ, Park HY (2017) An analysis of determinants for artificial intelligence industry competitiveness. *J Korea Inst Inf Commun Eng* 21(4):663–671
4. Xu Y, Helal A (2016) Scalable cloud-sensor architecture for the internet of things. *IEEE Internet Things J* 3(3)
5. Covington P, Adams J, Sargin E (2016) Deep neural networks for youtube recommendations. In: *Proceedings of the 10th ACM conference on recommender systems (RecSys '16)*. Association

- for Computing Machinery, New York, NY, USA, pp 191–198. <https://doi.org/10.1145/2959100.2959190>
6. Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of Things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comput Syst* 29(7):1645–1660
 7. Dolui K, Datta SK (2017) Comparison of edge computing implementations: for computing, cloudlet and mobile edge computing. In: *Global internet of things summit (GIoTS)*
 8. Xu LD, He W, Li S (2014) Internet of things in industries: a survey. *IEEE Trans Ind Inform* 10(4):2233–2243
 9. Song J, Lee B, Kim KT, Youn HY (2017) Expert system-based context awareness for edge computing in IoT environment. *J Internet Comput Serv (JICS)* 18(2):21–30
 10. Li H, Ota K, Dong M (2018) Learning IoT in edge: deep learning for the internet of things with edge computing. *IEEE Netw* 32(1):96–101 <https://doi.org/10.1109/MNET.2018.1700202>

Performance Analysis of Single-Pulse Modulation in Factory Environment Based on LiFi Standard



Ho Kyung Yu and Jeong Gon Kim

Abstract Internet of Things (IoT) technology is also widely used in industry. However, in recent years, various products tend to be mass-produced in a short period of time. Therefore, it is necessary to frequently change the location of the machine in the factory. In this situation, it is effective to communicate with wireless Light Fidelity (LiFi) instead of wired communication. In this paper, we compare the bit error rate (BER) and throughput by using Channel Impulse Response (CIR) between Light Emitting Diode (LED) and Photo Diode (PD) in the Industrial Wireless environments using On-Off Keying (OOK), 4-Pulse Amplitude Modulation (PAM), and 8-PAM in single carrier model. we use the frontend model filter for realistic implementation.

Keywords LiFi · VLC · Single carrier modulation

1 Introduction

With the development of Internet Of Things (IoT) technology, IoT technology is used in various fields such as life, commerce, and industry. In the recent industrial field, the period of the trend is shortened, and a product is produced in a short period of time in a large amount. And after the trend of the product, it became an industrial environment to produce other products. This environment is an optimal condition for using IoT technology.

In the conventional industrial sector, after the machine was installed, the product was produced by moving the product through the conveyor belt. And when the factory had to produce another product, it had to reposition the machine. In an environment where these machines are fixed and connected by wire, it takes a lot of money and time to install a new wired network while changing the position of the machine.

H. K. Yu · J. G. Kim (✉)

Department of Electronics Engineering, Korea Polytechnic University, 237 Sangidaehak-Ro, Siheung-Si, Gyeonggi-Do, Korea
e-mail: jgkim@kpu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_7

Thus, connecting the machines wirelessly can save these costs and allow flexible placement of machines to produce various products

In wireless communication, there are existing Radio Frequency (RF) communication and a new standard Light-Fidelity (LiFi) communication method. An industrial environment is an environment in which many machines communicate in a narrow space. When RF communication is used in this environment, severe interference occurs. And to reduce this interference, we need to build a radio map. Therefore, when the location of the machine is changed, the maintenance cost is newly incurred because the Access Point (AP) must be installed by constructing the radio map again.

However, using LiFi solves this problem. LiFi is a technology that uses a light-emitting diode (LED) and a photo diode (PD) to communicate using high frequencies in the visible light band. LiFi can communicate over a wide spectral range using visible light bands from 430 THz to 790 THz. Therefore, even if many machines communicate, they are not interfered. In addition, since the industrial environment is performed indoors, interference from external light is blocked. As a result, the reliability of communication is enhanced, and the internal light does not go outside, thereby enhancing security. This advantage makes effective use of IoT devices in industrial environments using LiFi [1].

IoT devices require the use of various modulation methods depending on two situations. The first is a situation in which a low reliability is required but a high-speed communication is required, and a second is a situation in which a reliable communication is required even if the speed is low. This communication speed and reliability should be adjusted according to the environment. The modulation schemes largely include single-carrier modulation and multi-carrier modulation.

Single carrier modulation requires high reliability even at low speeds. On-off keying (OOK), Pulse Amplitude Modulation (PAM), and Pulse Position Modulation (PPM) methods are available. Multicarrier modulation is relatively unreliable but is used when high-speed communication is required. Direct current-biased Optical Orthogonal Frequency Division Multiplexing (DCO-OFDM) and Asymmetrically Clipped Optical Orthogonal Frequency Division Multiplexing (ACO-OFDM) have. And the modulation method based on LiFi using visible light is Color Shift Keying (CSK) [2]. Reliable communication is more effective in industrial wireless environments. Therefore, single carrier modulation is used in this paper.

In this paper, the simulation is performed using the Channel Impulse Response (CIR) value in the industrial radio environment provided by the IEEE 802.11 TGbb. Using the frontend model filter, realistic simulation is implemented by further considering losses in LED and PD. The Frontend Model Filter implements drivers attached to the LEDs and PDs that serve as Tx and Rx, respectively. As the modulation method, we compare the change of BER and throughput according to each E_b/N_0 using OOK, 4-PAM and 8-PAM of Single Carrier Modulation. And it discusses how to use the LiFi effectively in the Industrial wireless environment.

In this paper, we describe the system model in Chap. 2 and proceed to Single Carrier Modulation Simulation in Chap. 3. The results and analysis will be presented in Sect. 4 while a conclusion will be given in Sect. 5.

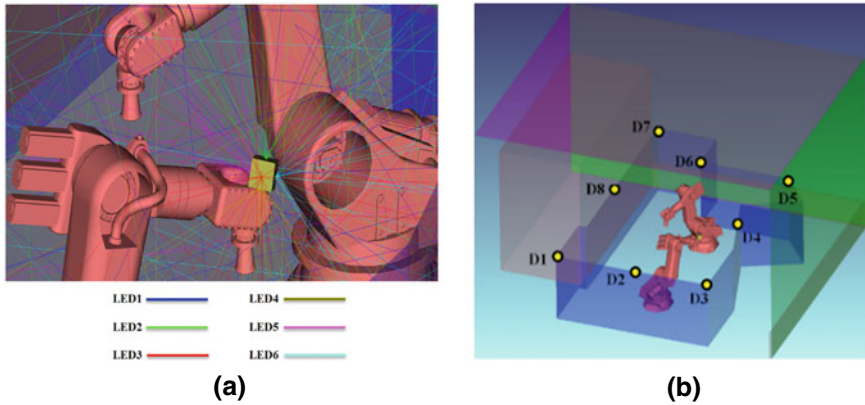


Fig. 1 a Location of LED in industrial wireless b location of PD in industrial wireless

2 System Model

In this paper, we use the simulation environment provided by IEEE TGbb to realistically realize the Industrial Wireless environment using VLC. Figure 1 shows two robots in one cell in the Industrial Wireless environment. The wall and floor are made of Concrete and the ceiling is made of aluminum metal. The size of the room with the machine is $8.03 \text{ m} \times 9.45 \text{ m} \times 6.8 \text{ m}$. The height of the robot is 2.7 m and the height of Plexiglas boundary is 2.5 m. Figure 1a and b shows the position of transmitter and receiver, respectively. The transmitter is placed in the shape of a cube, with the LEDs on each side. Six LEDs are composed of S1—S6. The half viewing angle and power per each luminaire are 60° and 1 cm^2 . In this paper, we use 6 LEDs to communicate. The receiver is attached to a simple wall 2.5 m high. The total number of receivers is 8, and each name is D1—D8. In this paper, we use D7 receiver. The FOV and the area of the detector are 60° and 1 cm^2 .

3 Single Carrier Modulation Simulation

In this paper, realistic VLC simulation using pulse modulation in Industrial Wireless environment was conducted through MATLAB. The pulse modulation process is shown as a block diagram as shown in Fig. 2. First, a random bit sequence is generated, and mapping is performed using the OOK scheme, the 4-PAM scheme, and the 8-PAM scheme. The OOK method is divided into 2 steps, 4-PAM is divided into 7 steps, and 8-PAM is divided into 15 steps. It then passes through the Tx Frontend Model Filter.

The optical frontend for LC imposes impairments, which have a non-negligible impact on the performance, on the signal. Hence, these effects must be modeled

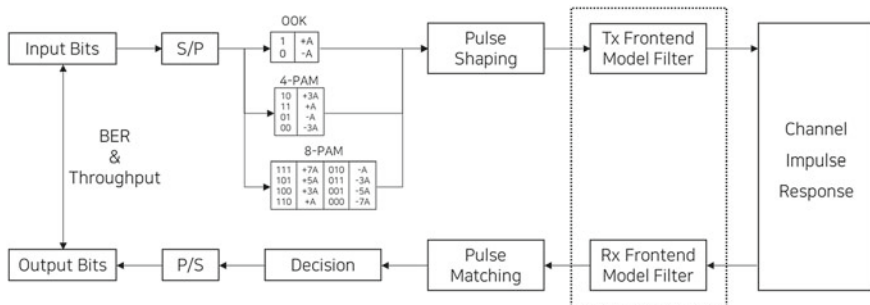


Fig. 2 Single carrier modulation block diagram

in addition to the propagation channel. The optical frontend model uses a highpass filter and a lowpass filter to create a filter model with MATLAB. The TX frontend comprises a driver electronics and a LED or laser diode. And the RX frontend comprises a photo diode and a bootstrap transimpedance amplified (TIA) [3].

After passing through the Tx Frontend model filter, the signal passes through the CIR provided by TGbb according to the simulation environment, and the equation of CIR is as follows.

$$h(t) = \sum_{i=1}^{N_r} P_i \delta(t - \tau_i) \quad (1)$$

$h(t)$ is the CIR value between the LED and the PD. Where P_i is the optical power of the i -th ray, τ_i is the propagation time of the i -th ray, $\delta(t)$ is the Dirac delta function and N_r is the number of rays received at the detector [4].

$$y(t) = h(t) \otimes x(t) + n(t) \quad (2)$$

Equation (2) is an equation to generate the output signal by convolving the original signal using CIR and then outputting the result. $y(t)$ is the output signal and $x(t)$ is the original signal. And $n(t)$ is AWGN and Noise Floor.

The signal passed through the CIR recovers the signal through the Rx frontend model filter. The recovered signal is demapped to determine the bit. The decoded signal is converted into a serial signal and compared with the original bit to calculate the BER value and throughput.

4 Results

In this paper, we have performed in the Industrial Wireless environment. The locations of Tx and Rx are fixed, and the main simulation parameters are summarized in Table 1.

Figure 3 shows the results of the simulation in the industrial wireless environment, in which data is transmitted using all six LEDs and received at D7. In the simulation, 6 LEDs each using 1 W were used. As a result, it was confirmed that the Eb/No value corresponding to the BER value of 10⁻⁵ was 91.1 dB for OOK, 99.3 dB for 4-PAM, and 102.3 dB for 8-PAM.

As a result of simulations, Eb/No, which is a BER of 10⁻⁵ required by data communication, was the lowest in OOK method and 8-PAM was the highest. This is because in the indoor LiFi communication simulation environment where the CIR value is low, the OOK method in which the power level of the signal is divided into

Table 1 Simulation parameter

Parameter	Value
Number of bits	1,000,000
Number of repeated counts	100
Bit time duration	100 ns
Bandwidth	10 MHz
Noise floor	-70dBm
Environment	Industrial Wireless
Point of Tx	All LEDs (S1-S6)
Point of Rx	D7
Optical CIRs	Overall—D7

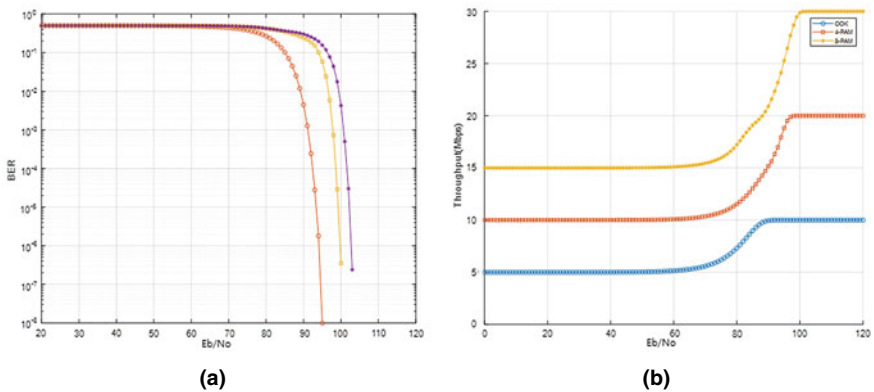


Fig. 3 a BER simulation results in industrial wireless b throughput simulation results in industrial wireless

two is low in the probability of occurrence of error due to the wide level interval. In the case of 8-PAM, It is possible to confirm that an error is likely to occur due to a large influence of noise.

The result of throughput is as follows. When OOK modulation method is used, it is confirmed that throughput is 10Mbps at 93 dB. 4-PAM requires 99 dB of Eb/No value to achieve throughput of 20Mbps. For 8-PAM, the throughput needs to be 102 dB for the Eb/No value to be 30Mbps. Throughput simulation results show that the throughput of 8-PAM is steadily high because 4-PAM sends 2 bits while OOK sends 1 bit, and 3-bit sends 8-PAM. However, when compared with the BER simulation results, 8-PAM requires higher Eb/No than OOK in order not to generate an error. Therefore, OOK should be used when reliable communication is required.

Through this paper, LiFi simulation in industrial wireless environment can confirm that OOK method is most effective for reliable communication considering BER and throughput.

5 Conclusion

In this paper, the simulation was performed using OOK, 4-PAM and 8-PAM pulse modulation in a realistic indoor LiFi environment. Simulation results show that Eb/No requiring BER value of 10^{-5} is influenced by LED power, number and modulation method. Depending on the modulation scheme, Eb/No required by OOK is the lowest, which is expected to be effective in environments requiring reliable communication like a factory. In the future, research will be conducted to compare with multi-carrier modulation schemes DCO-OFDM and ACO-OFDM

Acknowledgements This Work is supported by Individual Basic Research Program through Ministry of Education and National Research Foundation of Korea (NRF-2017R1D1A103035712)

References

1. Demers F, Yanikomeroğlu H, St-Hilaire M (2011) A survey of opportunities for free space optics in next generation cellular networks. In: 2011 9th Annual communication networks and services research conference, pp 210–216. IEEE, Ottawa
2. Islim MS, Haas H (2016) Modulation techniques for li-fi. *ZTE Commun* 14(2):29–40
3. Luo O, Chen J, Li J (2018) Modulation schemes for optical wireless communication. Doc: IEEE 802.11-18/0556r1. <https://mentor.ieee.org/802.11/dcn/18/11-18-0556-01-00lc-modulation-schemes-for-optical-wireless-communications.pptx>
4. Uysal M, Miramirkhani F, Baykas T, Qaraqe K (2018) IEEE 802.11bb reference channel models for indoor environments. Doc: IEEE 802.11-18/1582r4. <https://mentor.ieee.org/802.11/dcn/18/11-18-1582-04-00bb-ieee-802-11bb-reference-channel-models-for-indoor-environments.pdf>

Deep Learning-Based Experimentation for Predicting Secondary Structure of Amino Acid Sequence



Syntia Widyayuningtias Putri Listio, Ermal Elbasani, Tae-Jin Oh, Bongjae Kim, and Jeong-Dong Kim

Abstract The growth of biological databases has been increased in the past decades makes a protein structure prediction is one of the most important and challenging problems in Bioinformatics. The recent growth of Neural Network that have been shown promising result and also became indispensable tools in Bioinformatics. Using Convolutional Neural Network and Recurrent Neural Network with Long-Short Term Memory to predicting the secondary structure, our experiment shows a high result for RNN LSTM with accuracy 88.74% and loss rate 3.64%. In addition, for the CNN methods with an accuracy 87.74% and loss rate 3.85% prove that the latest technology of neural network also effective to be applied in secondary structure prediction of the proteins.

Keywords Protein secondary structure · Convolutional neural network · Recurrent neural network · Deep learning

S. W. P. Listio · E. Elbasani · B. Kim · J.-D. Kim (✉)
Department of Computer Science and Engineering, Sun Moon University, Asan, South Korea
e-mail: kjdvhu@gmail.com

S. W. P. Listio
e-mail: tia.listio@gmail.com

E. Elbasani
e-mail: ermal.elbasani@gmail.com

B. Kim
e-mail: bjkim0422@gmail.com

T.-J. Oh
Department of BT-Convergent Pharmaceutical Engineering, Sun Moon University, Asan, South Korea
e-mail: tjoh3782@sunmoon.ac.kr

T.-J. Oh · J.-D. Kim
Genome-Based BioIT Convergence Institute, Asan, South Korea

T.-J. Oh
Department of Pharmaceutical Engineering and Biotechnology, Sun Moon University, Asan, South Korea

1 Introduction

The development and rapid growth of numerous biological databases that store DNA and RNA sequence, protein sequence and other macromolecular structure data and make Neural Networks gradually became one of the tools in Bioinformatics. One of the popular challenges in Bioinformatics fields is to determine the structure of protein [1]. Proteins are macromolecules and have four different levels of structure that are primary, secondary, tertiary and quaternary. Secondary structure, refers to local folded structure that form within a polypeptide due to interaction between atoms of the backbone.

The mostly protein structure is largely determined by its primary structure that is simply the sequence of amino acids in a polypeptide chain [2]. In advanced studies show that accurate prediction of tertiary structures which is the overall three-dimensional structure of polypeptide was considered as a challenge but still with poor performance recently. The prediction of protein secondary structure from sequence is then considered as an intermediate problem bridging the gap between the primary sequences and tertiary structure prediction [3]. Protein secondary structure is the local three-dimensional (3D) organization of its peptide segments. There were two regular secondary structure states: Helix (H) and sheet (E), and one irregular secondary structure type: coil (C). In 1983 Sander [4] developed a secondary structure assignment method DSSP (Dictionary of Secondary Structure of Protein), which classified secondary structure into eight states ($H = \alpha$ -helix, $E =$ extended strand, $B =$ residues in isolated β -bridge, $G = 310$ -helix, $I = \pi$ -helix, $T =$ hydrogen bonded turn, $S =$ bend and $C =$ coil, the remaining). These eight states were often reduced to three states termed helix, sheet and coil respectively. The most widely used convention was that G , H and I were reduced to helix (H); B and E were reduced to sheet (E), and all other states were reduced to coil (C) [5–7].

In the challenge of predicting secondary structure of the protein there were many different algorithm and method, mostly of them were machine learning method such as Hidden Markov Model (HMM) [8] and support vector machines [9]. And also, there were many machine learning algorithms that can be applied to predicting the secondary structure of protein. This paper presented the comparison of predicting the secondary structure of protein using two neural network methods CNN, and RNN that has been showed a promising result to predicting protein secondary structure based on the sequence [10] and optimized it with LSTM. The purpose of this paper is to make modification on the hyperparameter of the network to prove that the modification on the experiment can optimize the result of the network for a better prediction.

2 Experimental and Datasets

CNN and RNN LSTM chosen as the prominent methods in Bioinformatics for predicting the secondary structure of the protein based on the sequence or primary structure. Using primary structure as an input the network train and test the date to make a highly accurate prediction of secondary structure of the protein. In this experiment, using the same data as an input the data trained using both of the models which developed in python and Keras library use similar code that available from [11].

3 Experimental Methods

In Fig. 1 shows the CNN model consists of 6 parts: embedding layer that will convert the dictionary integer input into float, three convolutional layers and also 2 dropout layouts with rate 0.3 between the convolutional layers.

For the RNN LSTM mode that we use for the experiment consists of 3 parts which is starting with embedding layers that have the same function to convert the input into float and then it's feeding the data into bidirectional layer to be trained and the last part is Time distributed dense layer as can be seen in Fig. 2.

3.1 Datasets

The data that used in this experiment, it is available for download from [7] that acquired from RSBC Protein Data Bank downloaded at 2018-06-06. The dataset has been transformed to complete the requirement of the network. Which are both eight state (Q8) and three states (Q3) structure sequence are listed. Also, all nonstandard amino acids, which includes B, O, U, X and Z are masked with "*" character. Also, there was an additional column to indicate whether the protein sequence contains nonstandard amino acids. The column description of the datasets explained at Table 1.

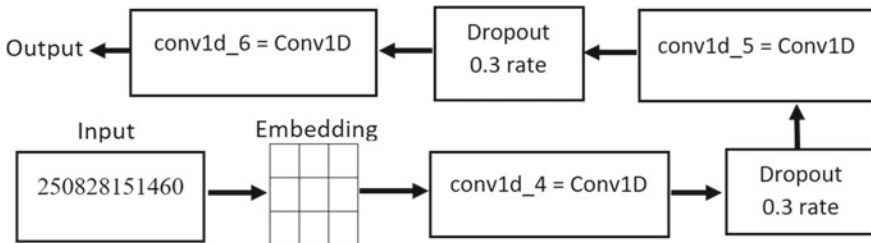


Fig. 1 Proposed CNN model

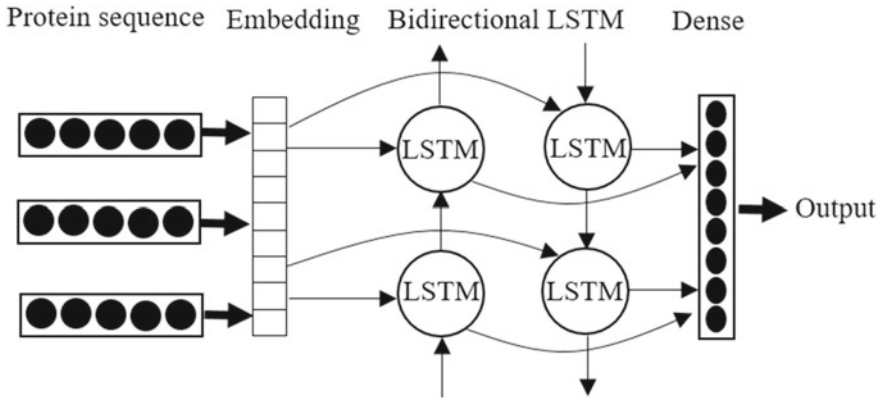


Fig. 2 Proposed RNN LSTM model

Table 1 Description of dataset for experiment

Column id	Description
pdb_id	The id used to locate its entry on protein data bank
chain_code	When a protein consists of multiple peptides (chains), the chain code is needed to locate particular one
seq	The sequence of the peptides
sst8	The eight-state (Q8) secondary structure
sst3	The three-state (Q3) secondary structure
len	The length of the peptide
has_nonstd_aa	Whether the peptide contains nonstandard amino acid (B, O, U, X or Z)

3.2 Experimental Result

In this experiment, using the same data as an input it trained using both of the models which developed in python programming language and Keras library. The proposed our model converged with only 30 epochs and learning rate at 0.001 with maximum length of the sequence that will be train is 40, shown in Figs. 3 and 4 is the graphic picture of the accuracy in training and test on using CNN and RNN LSTM methods. In the training phase the accuracy from the RNN LSTM achieve currency 87.42%, which was slightly higher than CNN (87.19%). Also, in the loss rate score, the performance of RNN LSTM method (3.79%) still show a better result than CNN method (3.87%) as can be seen in the Figs. 5 and 6.

To improve the performance of the network, modification with altering the hyper-parameter which is the learning rate, epoch and also the maximum length of the sequence. With increasing the epochs from 30 into 50, decrease the learning rate in 0.0005 and also increase the maximum length of the sequence from 40 into 70 so the variability of the input data also could be increased. As can be seen from the graphs

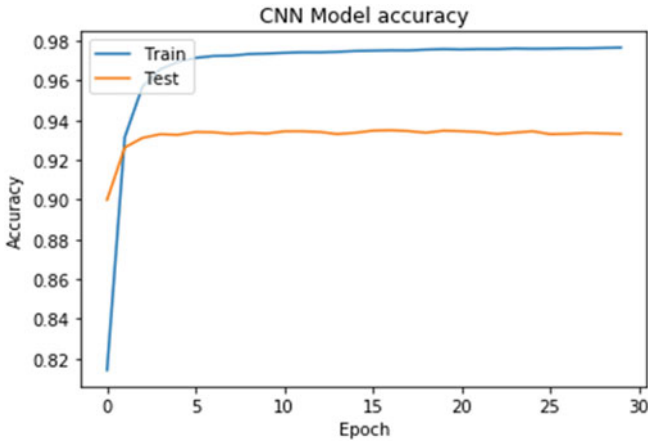


Fig. 3 The accuracy of CNN (epoch = 30, learning rate = 0.001)

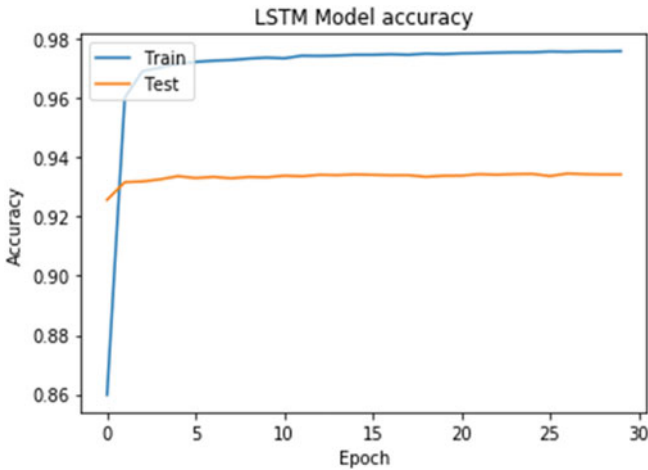


Fig. 4 The accuracy of RNN LSTM (epoch = 30, learning rate = 0.001)

in Figs. 7 and 8. The accuracy rate from both methods increase even though the RNN LSTM model (88.51%) still show a better result than CNN (87.74%). The similar result also shown on the Figs. 9 and 10 that present the loss rate of the methods.

4 Discussion

Deep Learning has been used in many fields in bioinformatic since its performance showed a promising result and one of them in the sequence predicting field. Based

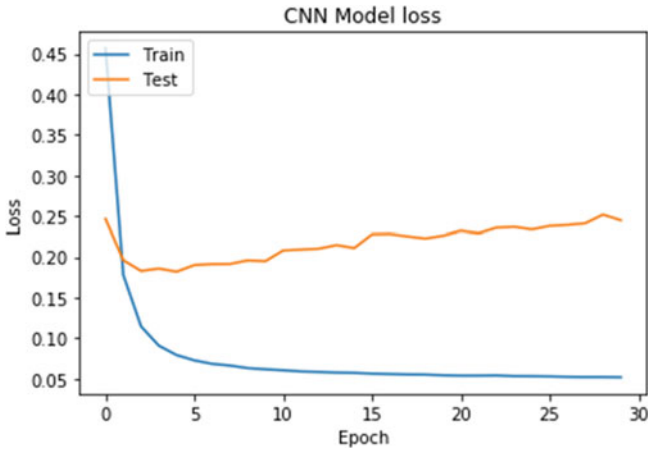


Fig. 5 The loss rate of CNN (epoch = 30, learning rate = 0.001)

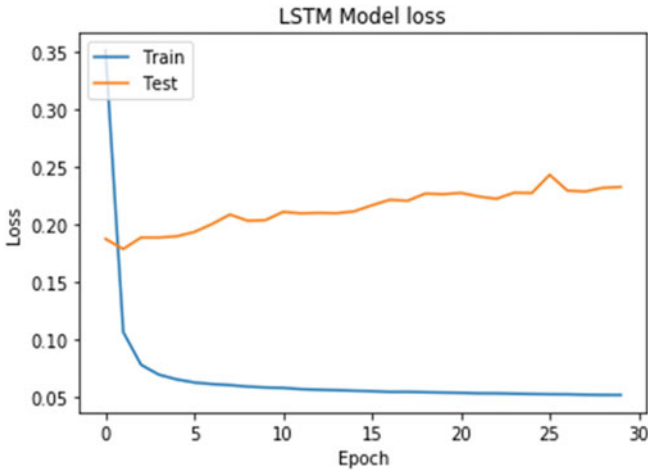


Fig. 6 The loss rate of RNN LSTM (epoch = 30, learning rate = 0.001)

on the experiment result on the both CNN and RNN LSTM methods showed that both deep learning methods achieve a good result both on the accuracy and loss rate. Furthermore, modification of the experiment with changes on the epoch, learning rate and also the maximum length of the sequence the network also shows an improvement on the accuracy rate. The growth of biological databases also could affect the result of the training and test using neural network, that's why this also could be a future work to find out the optimal number of epochs, learning rate and with the latest data to increase the performance of the deep learning methods to gain a better prediction.

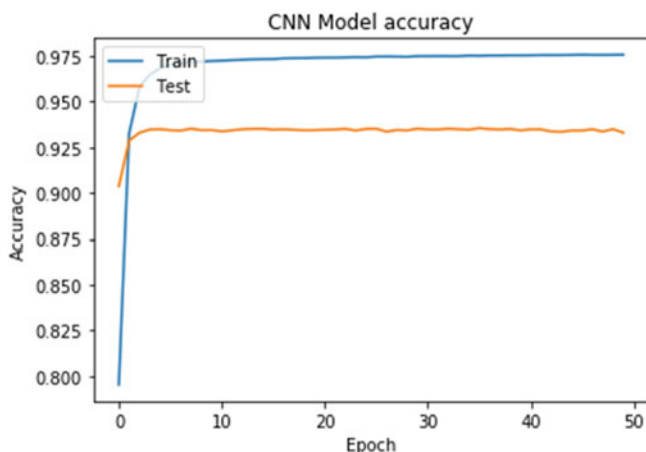


Fig. 7 The accuracy of CNN (epoch = 50, learning rate = 0.0005)

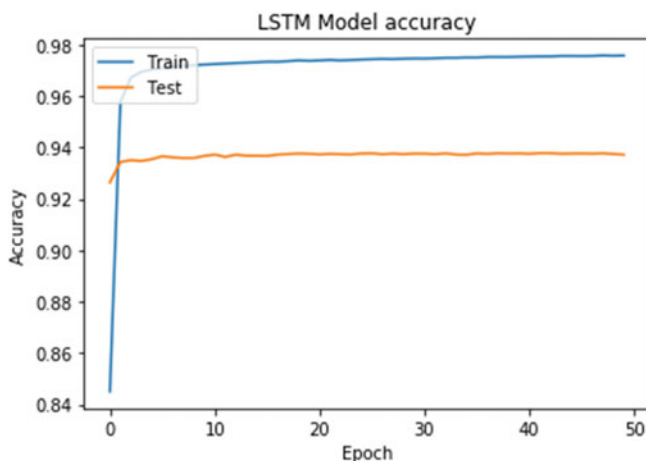


Fig. 8 The accuracy of RNN LSTM (epoch = 50, learning rate = 0.0005)

5 Conclusion

In this paper, deep learning methods CNN and RNN LSTM applied to make a prediction of secondary structure of the amino acids based on the sequence as the input. In this experiment with some data input RNN LSTM method showed better performance with an average training accuracy of 88.51% and loss rate 3.64%. Meanwhile, with small different CNN achieved training accuracy 87.74% and loss rate 3.85%. This result we accomplish with some modification on the number of the epoch, decrease the learning rate and increase the maximum length of the sequence. Based

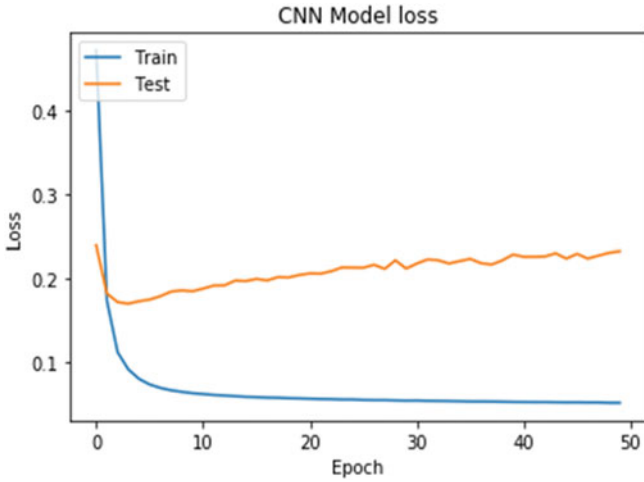


Fig. 9 The loss rate of CNN (epoch = 50, learning rate = 0.0005)

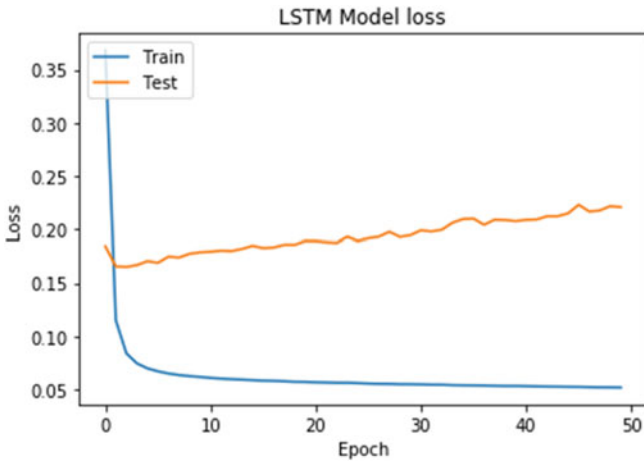


Fig. 10 The loss rate of RNN LSTM (epoch = 50, learning rate = 0.0005)

on this experiment we can see that deep learning will find a widely application in protein prediction on bioinformatics.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1058394), and the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW (2018-0-01865) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Chen K, Kurgan LA (2012) Neural networks in bioinformatics. Handbook of natural computing, pp 565–583. Springer, Berlin, Heidelberg
2. Anfinsen CB, Haber E, Sela M, White FH Jr (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A 47(9):1309–1314
3. Zhang B, Li J, Lü Q (2018) Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics. 19:293
4. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Res Biomolecules 22(12):2577–2637
5. Wang S, Peng J, Ma J, Xu J (2016) Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 6:18962
6. Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K et al (2016) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 19(3):482–494
7. Protein secondary structure (2019). <https://www.kaggle.com/alfrandom/protein-secondary-structure>
8. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics (Oxford, England) 14(10):846–856
9. Ward JJ, McGuffin LJ, Buxton BF, Jones DT (2003) Secondary structure prediction with support vector machines. Bioinformatics 19(13):1650–1655
10. Babaei S, Geranmayeh A, Seyyedsalehi SA (2010) Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. Comput Methods Program Biomed 100(3):237–247
11. Prostruct (2019). <https://github.com/vmalepati1/Prostruct>

A Study on Vulnerabilities of Linux Password and Countermeasures



Sanghun Kim and Taenam Cho

Abstract Linux is a representative operating system running on many servers and password is the primitive authentication way. In this paper, we analyze the vulnerabilities of Linux password file management and show that it is possible to use spoofed attack using system administrator privileges. We also suggests countermeasures to prevent this.

Keywords Linux · Password · Shadow file · Root privilege

1 Introduction

All computer systems use passwords as the most basic means of authenticating users. Linux is widely used by attackers because it is an open source operating system and widely available in various distributions. In order to protect the passwords of many users stored on Linux servers from attackers, the passwords are stored after encrypted using the one-way hash function. This makes it impossible for a system administrator, including root, to know other users' passwords. However, root has privileges including file access rights including password related files. If the root exploits this, he can disguise others to act maliciously and capture victims. This paper shows that this behavior is possible and suggests preventive measures.

S. Kim (✉) · T. Cho
Woosuk University, Jeollabuk-Do, Wanju-gun, South Korea
e-mail: schmid_t@naver.com

T. Cho
e-mail: tncho@ws.ac.kr

```

master@ITLinux:~$ ls -al /etc/passwd /etc/shadow
-rw-r--r-- 1 root root 3288 8월 11 01:04 /etc/passwd
-rw-r----- 1 root shadow 3434 8월 11 02:49 /etc/shadow

```

Fig. 1 Access rights for passwd and shadow files

```

victim:$6$3NTEu3ri$N9H1yS05Lzg84LRNLU6f0Jk4Rko
c8pmL8hH7MKY/VcrLffbwhQxvns8X7n5Arza7l63sjwQpZ
jQvANIDlhZ5G0:18118:0:99999:7:::

```

Fig. 2 Format of shadow file

2 Linux Password Management Scheme

Linux comes in many distributions, including Fedora, RedHat, CentOS, and Ubuntu. Since the core functions of distributions are the same, this paper is aimed at Ubuntu which is widely used.

2.1 Linux Password File Structure

Password related files are `/etc./passwd` and `/etc./shadow` files. The `passwd` file contains all user ids, passwords, home directories, and default shells. Only root has write permission to this file, but read access right is given to all users at risk. Therefore, Linux removes passwords from the `passwd` file and stores password-related information in the `shadow` file, allowing write access for this file only to root and no read access to others (Fig. 1) [1].

The structure of the shadow file is shown in Fig. 2 [2]. Each field is separated by a colon (:), where the second field is the information of user's password. This field consists of three sub-fields separated by \$. To prevent even root from getting user's password, one-way hash functions is applied to the password. The first sub-field is the hash function id used and the third sub-field is the hashed password. The second sub-field is a random number called "salt" generated by the system to prevent the dictionary attack. The password is concatenated with the salt before it is hashed. That is, `stored_password = hash_function (user's password || salt)`.

2.2 Scheme of Password Management

All users can change their own password using via `passwd` command [3]. But it requires users their current password. root has a privilege for users who have forgotten their password. However, since no one knows the hashed password, root can just

set the password with a new value using `passwd` command. If root changes the password without a user's consent to impersonate the user, the user recognizes it when he tried to login. root has almost all privileges because it must cope with system malfunction or user's mistake. By default, in order to protect a root account with powerful privileges, Ubuntu forbids root from remote login. However, any user who knows the root password can execute commands that require root privileges through the "sudo" command. In this paper, the user who knows the root password is referred to as root.

3 Impersonation Attack

3.1 Observation

As shown in Fig. 1, root has read and write access to the shadow file. So root can not only access the shadow file via `passwd`, but he can also copy, delete, overwrite the shadow file and modify it with an editor like `vi`. A hash function always produces the same value for the same input. Because of this property, system can authenticate a user by hashing the password entered by the user with the salt stored in the shadow file and comparing the result with the value of the shadow file.

3.2 Attack Scenario

The scenario was run on an "ITLinux" system on which Ubuntu-18.04 was installed with kernel-image-5.2.1, and the user id as root is "master".

- (1) master selects a target user "victim".
- (2) master copies the shadow file or copies the victim's password field from the shadow file (Fig. 3).
- (3) master sets the password of victim to a new password using the `passwd` command (Fig. 4). In our experiments, the old password was "victim" and the new one was "not".
- (4) master logs in as the victim using the new password and perform the malicious task as he wants (Fig. 5).
- (5) master logs in with his id and restores the shadow file with the copied file or modifies the shadow file with the copied previous password field of the victim (Fig. 6).

```
master@ITLinux:~$ sudo cp /etc/shadow shadow
```

Fig. 3 Backup of the original shadow file

```

master@ITLinux:~$ sudo passwd victim
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully

```

Fig. 4 Reset of victim's password

```

login as: victim
victim@114.71.84.214's password:
Welcome to Ubuntu 18.04.2 LTS (GNU/Linux 4.15.0-54-gen
eric x86_64)
victim@ITLinux:~$ cat message.txt
This is a message from the compromised administrator!!
victim@ITLinux:~$ █

```

Fig. 5 Impersonated behavior

```

master@ITLinux:~$ sudo mv shadow /etc/shadow

```

Fig. 6 Restore of shadow file

3.3 Results of Experiments

After master restores the shadow file, we confirm that the user victim can log in with the previous password, "victim". (Figure 7).

```

login as: victim
victim@114.71.84.214's password:
Welcome to Ubuntu 18.04.2 LTS (GNU/Linux 4.15.0-54-gen
eric x86_64)
victim@ITLinux:~$ cat message.txt
This is a message from the compromised administrator!!
victim@ITLinux:~$ █

```

Fig. 7 Login using victim's original password

4 Countermeasures

4.1 Key Ideas

The above scenario is possible due to the vulnerability that root can modify the `/etc/shadow` file with not only `passwd` but also other commands. To compensate for this vulnerability, we modified the kernel so that even the root cannot modify the `/etc/shadow` file in any way except by `passwd` command.

4.2 Requirements

The modified kernel should support the following:

- (1) No one including root can modify shadow file with editors such as `vi`.
- (2) No one including root can overwrite shadow file with other files.
- (3) No one including root can access illegally with any relative path name.
- (4) Any other file access control remains as it is.

4.3 Implementation

We modified Kernel-image-5.2.1 in Ubuntu-18.04 to meet the requirements, ported it to the server named “test”, and then tested whether the requirements were met. Whenever commands to access files are executed, `do_sys_open()` in `openat()` is called. Access is controlled using information about the file to be accessed as shown in Fig. 8. The functions modified or added are shown in Table 1.

4.4 Result of Experiments

We tested whether the modified system meets the requirements. The administrator id for the root role in our experiment is “newmaster”.

- (1) Modifying shadow files with a editor

It failed when root attempt to save the shadow file with `vi` editor (Fig. 9).

- (2) Replace with shadow_backup (the copied shadow file) `/etc/shadow`

Attempting to overwrite the backed up shadow file (`shadow_backup`) to `/etc/shadow` failed (Fig. 10). When he tried it with relative path, it also failed.

- (3) Changing password using `passwd` command

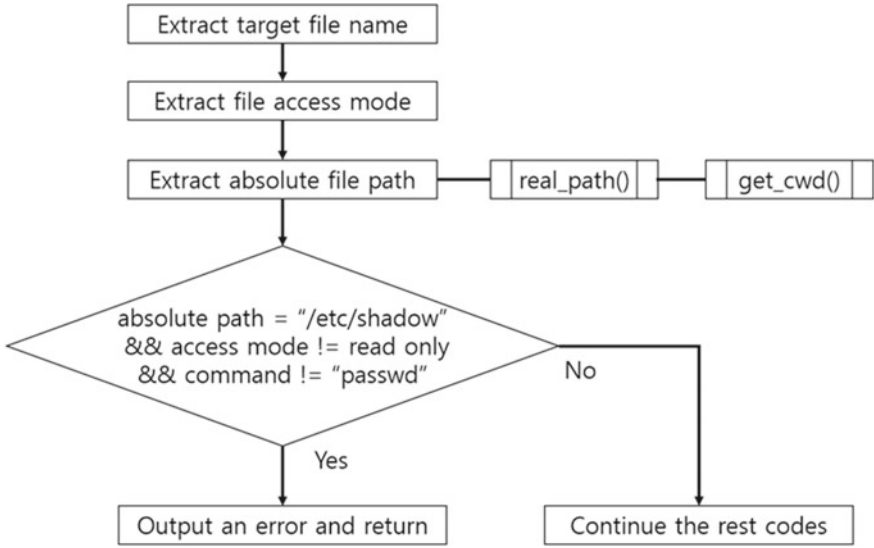


Fig. 8 Flowchart of modified do_sys_open() code

Table 1 Related System Functions

Name	Function
opentat()	Accept user’s file related command
do_sys_open()	Check target file name and access mode
real_path()	Extract Absolute path
get_cwd()	Extract current working directory

```

"shadow"
"shadow" E212: Can't open file for writing
Press ENTER or type command to continue
  
```

Fig. 9 Fail of modification using vi

```

test@test:~$ sudo cp /etc/shadow shadow
test@test:~$ sudo mv shadow /etc/shadow
mv: cannot move 'shadow' to '/etc/shadow': Permission denied
  
```

Fig. 10 Fail of overwriting using mv

Changing the password of the victim using the passwd command by root or victim worked fine. (Figure 11).

Fig. 11 Renew password using passwd

```
test@test:~$ sudo passwd victim
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
```

5 Conclusions and Future Works

In this paper, we show that it is possible to exploit the privilege granted to root to manage Linux system and change a user’s password and disguise it as the user. Also, to prevent this, we modified file related system functions in Linux kernel to prevent root from abnormal access for the shadow file. In the future, we will study for vulnerabilities using other commands or ways to access shadow file such as chpasswd and symbolic /hard links. We will also study how to use logs related to root’s abnormal behavior for the shadow file.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03032637).

References

1. GitHub manual. <https://github.com/shadow-maint/shadow>, GitHub
2. GitHub, File formats and conversions-SHADOW. <http://man7.org/linux/man-pages/man5/shadow.5.html>, GitHub
3. GitHub, user command-PASSWD. <http://man7.org/linux/man-pages/man1/passwd.1.html>, GitHub

Analysis of Learning Model for Improvement of Software Education in Korea



Ji-Hoon Seo and Kil-Hong Joo

Abstract The importance of domestic software education is emphasized and the implementation of mandatory education is gradually spreading. Therefore, in this paper, the overall current state of software education in South Korea and five examples of teaching-learning methods that are currently applied were analyzed and problems in the teaching-learning methods were presented. Unstructured data were collected mainly from domestic software education related web sites to extract elements necessary in software education thought by teachers and learners, and based on the result, improvement points of five learning models that are currently used in South Korea were derived.

Keywords Software education · Data analysis · Computational thinking · Big data

1 Introduction

Civilization and science are advancing exponentially thanks to the creative efforts of mankind. Consequently, humans are looking forward to the age of the fourth industrial revolution beyond the third wave of Alvin Toffler. The Fourth Industrial Revolution is a software revolution that can be the beginning of a hyper-connected society based on IT, and accordingly, a new future society is anticipated. Predicting such changes in the environment, new educational model crazes are breaking out in educational circles all over the world. Thus far, domestic and foreign education has been using cramming education, which is gathering many students into each classroom and inducing learning, focusing on finding employment, from the 19th century in North America and Europe [1]. The cramming education and environment as such

J.-H. Seo

Incheon University, 119, Academy-ro Yeonsu-gu, Incheon, Republic of Korea

e-mail: jihoon@inu.ac.kr

K.-H. Joo (✉)

Gyeongin National University of Education, 115, Sammak-ro, Manan-gu, Anyang-si,

Gyeonggi-do, Republic of Korea

e-mail: khjoo@ginue.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_10

had been applied as they are even in the 21st century until they began to be changed little by little only very recently. Currently, in South Korea, rather than cramming education, creative education intended to derive learners' creative ideas and improve problem solving abilities is being implemented, and the national education system is also being changed based on the introduction of the free semester system intended to encourage learners' aptitudes and originality, and communication and cooperation centered convergence talent cultivation [2]. In addition, the education that is becoming a big issue along with the fourth industrial revolution is SW education. Thanks to the opinions that SW education is essential to be prepared for the future and that SW education should be implemented early centered on many leaders and influential entrepreneurs in the world, SW education is actively implemented in South Korea. However, the introduction of SW education into South Korea exhibits many problems despite that SW education is indispensable in future. In this paper, such problems were sought and a Korean style SW education class model suitable for South Korea will be presented centering on statistical cases of big data.

2 Related Works

2.1 Current State of Global Software Education

In the future information society, software-related capabilities will be regarded as very important due to diverse occupations and changes in jobs. Accordingly, various countries have been preparing diverse educational systems and environments to cultivate software-related capabilities [3] (Table 1).

The UK has been implementing a national curriculum that includes computing education as a required course in elementary and secondary schools since September 2014 and the United States has also announced a plan to carry forward computer science education policies for all elementary and middle school students from January 2016. Major countries are actively introducing software education, computer science, or computing including algorithms and programming into curricula as a required

Table 1 Current states of software education by country

Year	Implementing country	Current states of SW education
2014	England	Operates the subject "Computing" as a required subjects for students aged 5–16 years
2015	Estonia	All elementary schools conduct computing education
2016	Finland	Added 'ICT' to elementary, middle, and high school to educate on the principle of algorithm and coding
2016	United states	Determined to conduct AP course, the subject "computational thinking"

subject. Therefore, it can be predicted that software education will become an important measure for national competitiveness in the future [4].

2.2 Current State of Software Education in Korea

In response to the rapidly changing trends of global education, the South Korean government held a software-oriented society realizing strategy briefing session centering on four departments in July 2014 to begin discussion on software education in earnest through the announcement of plans to strengthen elementary/middle school SW education by the Ministry of Education. Thereafter, the South Korean government announced the “Plan to Carry Forward Talent Cultivation for Software-Oriented Society” in July 2015. In order to prepare a foundation according to the plan, on-the-job education was conducted for 60,000 elementary school teachers accounting for 30% of all elementary school teachers by 2018. Intensive training was carried out for 6,000 teachers among them. Intensive training was promoted for the entire 1,800 middle school ‘information’ subject teachers and teachers who had ‘information computer’ licenses. From 2018, software education was applied to elementary/middle/high school curricula in stages. The primary goal of the foregoing is to cultivate ‘futuristic creative talents’ equipped with problem-solving abilities who can implement their creative ideas with software by 2020 through software education for elementary/middle school students in South Korea [5].

2.3 Connectivity of Domestic Software Education

In South Korea, after the revision of curricula in 2015, all elementary school students and middle school students commonly became to be able to complete software education and after entering high school, the students were allowed to complete enrichment learning connected to careers (Table 2).

The software education presented in elementary school is organized in the practical course [6]. The subject consists of play centered activities intended to enhance learners’ understanding and provide fun and was made to enable students to easily and interestingly learn problem-solving methods centering on experience through the Educational Programming Language (EPL).

3 Proposed Method

Although the terms that refer to software education vary slightly from country to country, they are generally used together with terms such as computing education, computer science education, algorithm education, programming education, coding

Table 2 Connectivity of domestic software education

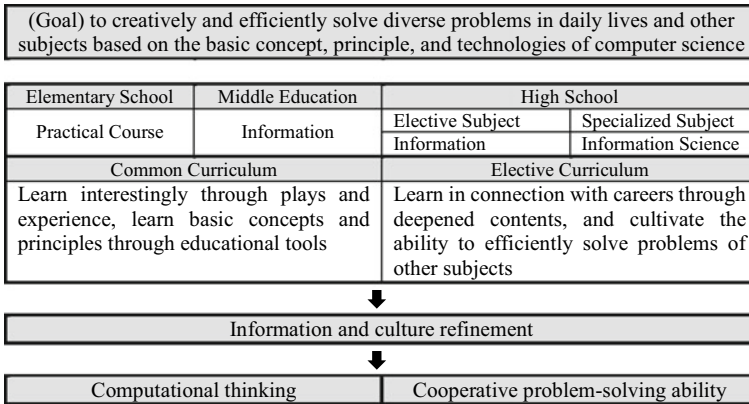


Table 3 Domestic software education learning intended to enhance computational thinking

Class model	Learning type
Demonstration centered model (DMM)	Demonstration-modeling-making
construction centered model (UMC)	Use-modify-recreate
Development centered model (DDD)	Discovery-design-development
Design centered model (NDIS)	Needs-design-implementation-share
CT element centered model (DPAAP)	Deconstruction-pattern recognition-abstraction-algorithm-programming

education, and information science education. This paper will present such problems and improve the models to present customized learning models suitable for Korean style software education (Table 3).

The five types of learning models are oriented toward domestic software education and composed to fir diverse environments. However, according to cases of application of these models, whereas there are cases where learning types were successfully improved, there are some other cases where these models were not suitable for Korean style software education. This paper will present such problems and improve the models to present customized learning models suitable for Korean style software education.

3.1 Analysis of the D-D-M Demonstration Centered Model

The demonstration centered model is a direct teaching model technique, which is a teacher-centered learning method that explains new technologies and concepts to large groups of learners in the stages of task assignment, responsibility assignment,

imitation, practice, and completion and assigns training according to the instruction of the teacher. Learning methods as such are useful when teaching EPL(education programming language) such as scratch and entry for the first time. This learning method is efficient for learning the functions of the educational software, but it just improve program execution speed and proficiency and not suitable for inducing problem solving methods and creativity. Since software education is not to learn how to use programs that have been already made but to learn the thinking process to solve problems creatively using the basic concepts and principles of computer science and develops the capability to solve problems based on computational thinking, education using this learning method is highly likely to become cramming education.

3.2 Analysis of the Reconstruction Centered Model

The reconstruction centered model is a method in which the learners first carry out an experience learning play and reconstruct the play into their programs while undergoing stages to modify and revise the play. This is a program in which the teacher intentionally modifies existing modules and algorithms and presents the outcomes to learners so that learners will reconstruct existing learning activities. Although this learning method can give curiosity and interest to students, the diversity of learning plays is important to graft it onto software education. However, it is not easy to connect plays under diverse themes to software education thereby grafting them onto software education.

3.3 Analysis of the Development Centered Model

The development centered model is a model for understanding of entire process of software development in terms of software engineering. The learning type was constructed based on the overall research process and design process for the programs that individual learners want to develop. This model is divided into three stages of research, design, and development. It induces research into algorithms through diverse unplugged activities for subjects that trigger interest to learners and carries out design based on resultant stakeholders' requirements thereby conducting coding education. Although this learning method is affective for getting jobs in South Korea and cultivating development talents, it is difficult to learn the stages to research into IoT technology to be prepared for the fourth industrial revolution and other physical computing, decompose algorithms, recognize patterns, and conduct abstraction.

3.4 Design-Centered Model

This learning method is a class model in which students experience and participate in the entire processes ranging from selecting a topic with a critical mind to survey, data collection, study, presentation, and evaluation based on the project class model. Teachers present only advices or simple feedback on diverse problems faced by students, and play only the role of helpers to help the learners derive meaningful results by themselves. This model can enhance the creativity of learners and it is a learning method widely used at universities both at home and abroad based on requirements analysis, creative design, development tool centered implementation, and sharing. This model, however, is useful for application to learners skilled in coding education to some extent and give interest to initial learners but does not provide a personal customized solution.

3.5 CT-Centered Model

The CT-centered model is implemented through the stages of decomposition, pattern recognition, abstraction, algorithm, and programming. This model presents cases for problem situations centering on the problem—solving learning method, identify resultant goals, and sets hypotheses to clearly identify problems thereby presenting solutions. The solutions are based on the process to explore problems, problem solving activities, and methods to understand the principles to learn the strategies. Since this method is based on the premise of KS3 and the four-stage module strategy presented by Google, clear components by stage according to CT are not defined in South Korea and learning methods that decompose the four-stage strategy to increase stages are also presented. Therefore, to graft this method onto Korean style software education, transparent clear stages should be set.

4 Analysis and Improvement Plan

In this paper, centering on the five types of learning models for software education being utilized as such, improvement plans will be presented through analysis of learner questionnaire data collected from the web. The following analysis figure shows words with features of software education extracted by collecting unstructured data for one year from the time point at which software education became mandatory centering on the web sites related to domestic software education and classifying learner data and teacher data using the clustering technique (Fig. 1).

From among certain words, sentiment based words were collected in order of frequencies beginning from those with the highest frequency in order to analyze the importance of software education thought by learners and that thought by teachers,

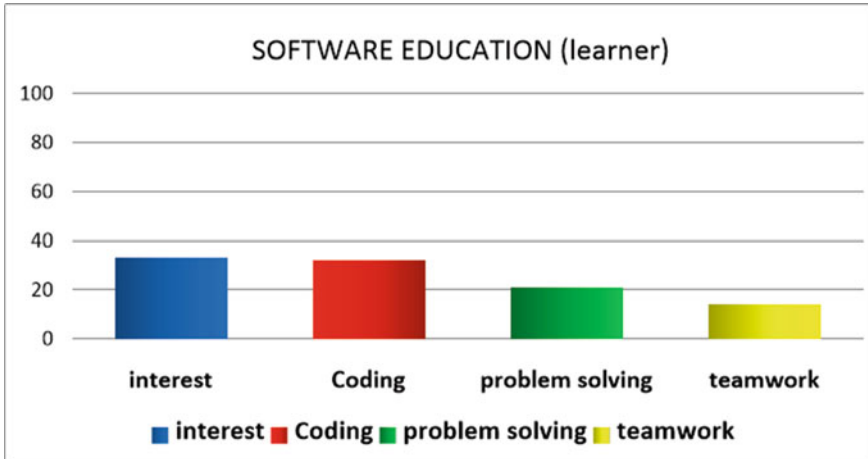


Fig. 1 Element necessary in software education thought by learners

which are necessary in this paper. Accordingly, the words thought to be the most important by teachers and learners were derived through analysis centering on the words, interest, the importance of coding, the importance of problem solving and teamwork among the most important elements in software education.

The first item is the elements necessary in software education thought by learners. The learners selected fun and interest as the elements thought to be the most important and presented coding learning method, problem solving ability, and teamwork as the next important elements. This indicates that rather than the importance of problem solving ability to carry out programming, learners first preferred leading effective learning through interest-oriented learning (Fig. 2).

The second analysis figure shows the analysis of the elements necessary in software education thought by teachers on the contrary. According to the results of the analysis, teachers greatly emphasized the problem-solving ability pursued by the advanced education, and presented coding ability, interest, and teamwork as next important elements. As shown in these results, different opinions appeared from the viewpoint of learners who receive software education and from the viewpoint of teachers who teach. Therefore, it can be recognized that the educational philosophies pursued by learners and teachers are different from each other in terms of education.

Through these results, it can be seen that to improve the five types of class models in South Korea, an effective method is first presenting curricula centering on learners' position. Although the problem solving ability thought to be important by teachers cannot be excluded because it is also important, the decomposition and abstraction model is still too difficult for young elementary school students to implement. Therefore, as the first improvement, reconstruction centered models should be first implemented by young learners in elementary education; provided that the learners should be led to encounter EPL in the demonstration centered model to construct demonstration learning methods based on convergence type plays. In addition, in the

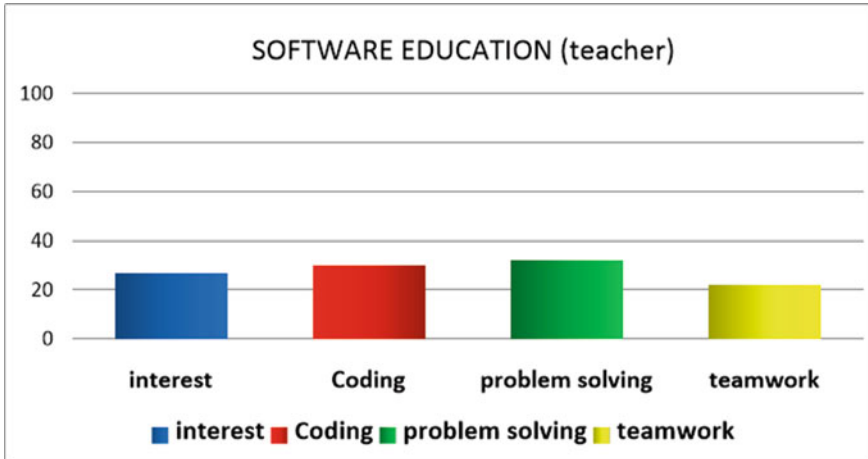


Fig. 2 Element necessary in software education thought by teachers

case of middle and high school students, excellent developers should be cultivated through curricula that actually emphasize teamwork and problem solving based on the stream of domestic IT. Therefore, the CT—centered model should be deepened to construct models than enable learning; provided that, the project technique of the design centered model should be fused with the forgoing model to construct personal customized curricula to be prepared for the future rather than interest for efficiency.

5 Conclusion

In this study, problems in Korean style software education were identified centering on five types of learning models, and education based unstructured were collected from the web to classify, compare, and analyze the important elements in software education thought by learners and teachers. Although the educational curricula pursued by learners and those pursued by teachers were derived differently in the results, if the five types of learning models are fused together centering the opinions as such considering learning ages, more effective Korean style software education methodologies can be obtained.

References

1. Won Joo Lee (2019) A study on software education donation model for the social care Class. *J Korea Soc Comput Inf* 24(1):239–246
2. Khan IA, Choi JT (2014) An application of educational data mining (EDM) technique for scholarship prediction. *Int J Softw Eng Appl* 8(12):31–42

3. Jun Woochun (2009) The current status of ICT education in Korea: teacher education perspective. *Asian J Teach Educ* 1(1):23–32
4. Park JH, Park S, Seo YG (2018) Software education and business model for enforcing the computational thinking ability. *Korea Digital Contents Society* 19(12):2297–2304
5. Seo J-H, Joo K-H, Park N-H (2018) A study on the decision-making of effective S/W education based on opinion mining analysis. *Int J Eng Technol* 7(4.16):5–9
6. Seo Ji-Hoon, Joo Kil-Hong (2018) Analysis of the elements of future development of korean style software education through the opinion mining technique. *Lect Notes Electr Eng* 474:1410–1415

Implementation and Experiment of Join Optimization Algorithm for Inverted Index in an RDBMS



Yoonmi Shin, Odsuren Temuujin, Minhyuk Jeon, Jinhyun Ahn,
and Dong-Hyuk Im

Abstract In a relational database management system (RDBMS), when a user searches for a keyword included in a document, a keyword search is attempted using various join methods. When a keyword search is performed using a merge join in an RDBMS that stores a large number of documents, the retrieval time for the query is increased due to unnecessary comparison operations. In this study, we propose a keyword search using a Skip Merge Join that minimizes unnecessary comparison operations when a keyword search is performed using a row direction relational inverse index table. We confirmed the results of improving the keyword search performance by implementing the Skip Merge Join algorithm using the relational database PostgreSQL and verifying it through experiments.

Keywords Keyword search · Join processing · Inverted index · Documents · Aggregates · Relational database · SQL

1 Introduction

The demand for big data processing is increasing rapidly due to the increase in data utilization. Therefore, text search support for big data is a major issue in relational database management systems (RDBMSs). Recently, a special Information

Y. Shin · O. Temuujin · M. Jeon · D.-H. Im (✉)
Department of Computer Engineering, Hoseo University, Asan, South Korea
e-mail: dhim@hoseo.edu

Y. Shin
e-mail: sinyoonmi12@gmail.com

O. Temuujin
e-mail: temuujintemka@gmail.com

M. Jeon
e-mail: jeoncoder@gmail.com

J. Ahn
Department of Management Information Systems, Jeju National University, Jeju, South Korea
e-mail: jha@jejunu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021
J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_11

retrieval (IR) engine has been integrated into the RDBMSs to optimize query execution and storage for text retrieval [1]. Text search for big data mainly uses complex query combinations and multiple searches within a relational database. Because these queries can take a considerable amount of time, the search system uses the index to increase its effectiveness. Commonly used index-based wide-range keyword searches use merge join to produce valid results in most cases. However, when using a merge join in the index, unnecessary comparison operations increase the retrieval time of the query.

ZigZag merge join [2] have been studied to reduce the search time of traditional merge joins in RDBMSs. The ZigZag merge join minimizes unnecessary comparison operations using a Gallop search [3] and shortens the search time. However, when a keyword search is performed using a Gallop search in a ZigZag merge join, if the comparison value is overpassed, the cursor is moved to find the comparison value using a binary search.

In this study, we propose an efficient keyword search method that can be applied to large document sets stored in database to solve these unnecessary cursor movements. By using the aggregate functions of the database, we can skip the correct number of rows and minimize unnecessary cursor movement. Therefore, it is shown that the merge join using exact skipping not only shortens the retrieval time but also reduces the number of cursor movements.

2 Related Work

Keyword searches in databases have been regularly studied [4–8]. BANKS [4] is a system for performing keyword-based retrieval in relational databases. The study in [4] proposed a heuristic algorithm for locating and ranking query results and was designed to avoid unnecessary tuple tree creation, while improving space complexity. The result of a query consists of a root tree that links the tuples matching the individual keywords in the query. The study in [5] proposed that compute all the interconnected tuple structures for a given keyword query using SQL. There are two steps. The reduction step remove the tuples that do not participate in any results using SQL. And

the join step handles relational algebraic representation using SQL on the reduced relation. The study results showed that efficiency was improved by the new tuple reduction approach, effectively eliminating unnecessary tuples in relationships and reducing the relationship to handle the final result.

3 Skip Merge Join

In this study, we propose two queries divided into conjunctive and phrase queries.

Algorithm 1. Skip Merge Join algorithm for conjunctive query

```

Begin
  Set coudor1 at beginning of row1
  Set coudor2 at beginning of row2
  While row1 is not null and row2 is not null loop
    If row1.docid = row2.docid then
      Return row2
      Move coudor1 to the next row of row1
      Move coudor2 to the next row of row2
    Elsif row1.docid < row2.docid then
      countTemp = SkipRow (row1.term,row1.docid,row2.docid)
      Move cursor1 by the countTemp of row 1
    Else
      countTemp = SkipRow (row2.term,row2.docid,row1.docid)
      Move cursor2 by the countTemp of row 2
    End if
  End loop
End

```

A conjunctive query is a query that finds a document containing n searched keywords. For example, if a user were to search for the two keywords “apple” and “fruit,” the query will find a document containing the two keywords. Skip Merge Join algorithms do not perform unnecessary comparison operations by skipping document ID smaller than the comparison value. This Skip Merge Join process uses aggregate functions. When the aggregate function is used, the cursor is moved to the correct position by determining the precise interval at which the cursor should move. The pseudo-code for the conjunctive queries is shown in Algorithm 1.

Algorithm 2. Skip Merge Join algorithm for phrase query

```

Begin
  Set coustor1 at beginning of row1
  Set coustor2 at beginning of row2
  While row1 is not null and row2 is not null loop
    If row1.docid = row2.docid then
      If row1.offset = row2.offset-1 then
        Return row2
        Move coustor1 to the next row of row1
        Move coustor2 to the next row of row2
      Elsif row1.offset < row2.offset-1 then
        countTemp = SkipRow(row1.term,row1.docid,row1.offset,row2.offset-1)
        Move cursor1 by the countTemp of row 1
      Else
        countTemp = SkipRow (row2.term,row2.docid,row2.offset,row2.offset)
        Move cursor2 by the countTemp of row 2
      Elsif row1.docid < row2.docid then
        countTemp = SkipRow (row1.term,row1.docid,row2.docid)
        Move cursor1 by the countTemp of row 1
      Else
        countTemp = SkipRow (row2.term,row2.docid,row1.docid)
        Move cursor2 by the countTemp of row 2
    End if
  End loop
End

```

Phrase queries in pseudo-code form in Algorithm 2 are extended types in conjunctive queries. A phrase queries is a query that finds documents with n keywords in succession. In [9], a restricted multi-predicate merge join algorithm was added to a Skip Merge Join. This query first moves the cursor to the document ID. If a discrepancy exists between two document IDs, the aggregate function is used to move the cursor to a document ID greater than or equal to the comparison document ID. If the document ID matches, then the offset of the keyword is compared. At this time, the keyword offset cursor movement uses the same method as that of the document ID.

The data storage method uses a row-oriented relational inverse index table. The main table consists of a term column and a docid column indicating the document that contains the term. The offset table contains the offset column, which indicates the term in the document. The main table is used for conjunctive queries, and the offsets table is used for phrase queries.

4 Experimental Results

In this study, we implemented ZigZag and Skip Merge Joins algorithms as functions using the relational database PostgreSQL. The query speed of 1–5 kb documents was compared with 50,000 and one million documents. The data from the inverted index table was used by the Westbury Lab [10] of USENET with 50,000 and one million documents out of approximately 21 million documents. For each document, inverted index data were generated by stop-word and stemming processing, which are natural language processing methods.

The performance test of the conjunctive query was compared with data from 50,000 and one million documents. First, two search keywords in a document were randomly selected and established as one set, and then the query speed and number of comparison results of 100 keyword sets were averaged.

The performance test of the phrase queries was compared with data from 50,000 and one million documents. The performance test of the phrase query selected 100 search keyword sets at random. We then produced a set of keywords in which one keyword is grouped with a subsequent keyword based on the selected keyword. In this manner, we averaged the execution time of the phrase search query and the number of comparisons in 50,000 and one million documents for the set of two keywords.

Figure 1 shows the results of the conjunctive and phrase queries. Specifically, it shows the query execution time and number of comparisons of 100 keyword sets in 50,000 and one million documents. According to the performance test results, Skip

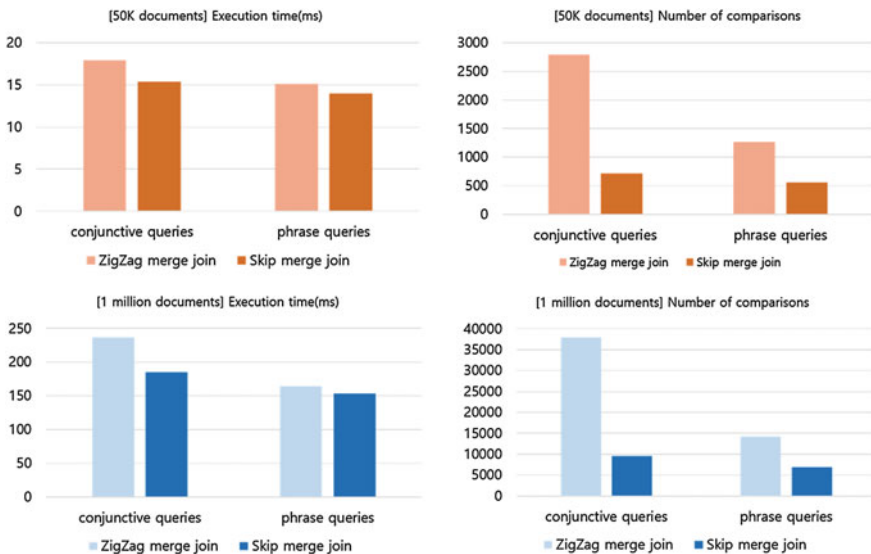


Fig. 1 Results of conjunctive and phrase query performances from experiments on 50,000 and 1 million documents. The left graph shows the query execution time; the right graph shows the number of comparisons

Merge Joins perform better than ZigZag merge joins in a set of two keywords. In addition, we confirmed that the query speed with one million queries was relatively reduced compared to that with 50,000 documents.

5 Conclusion

In this study, we proposed a Skip Merge Join algorithm to store large documents in a database and effectively process keyword searches of documents. Experimental results showed that both the query speed is fast and number of comparisons are reduced by using Skip Merge Join with an aggregate function of a database in conjunctive and phrase queries. In addition, the higher the number of documents in conjunctive and phrase query experiments, the faster the execution time in the use of Skip Merge Joins. Therefore, Skip Merge joins can yield better query performance than can ZigZag merge joins when extensive keyword searches are performed in large documents.

A future study will add a paragraph column that can be saved as a paragraph of documents to enable faster searches in which unnecessary offset rows are skipped during phrase queries. We will also explore means of saving space using partitioning techniques.

Acknowledgements This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant NRF-2017R1C1B1003600, in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program Supervised by the Institute for Information & Communications Technology. Promotion (IITP), under Grant IITP-2019-2018-0-01417, and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A1B07048380.

References

1. Hamilton J, Nayak T (2001) Microsoft sql server full-text search. *IEEE Data Eng Bull* 24(4)
2. Rae I, Halverson A, Naughton J (2014) In-RDBMS inverted indexes revisited. *ICDE 2014*, pp 352–363
3. Bentley JL, Yao AC-C (1976) An almost optimal algorithm for unbounded searching. *Inf Proc Lett* 5(3):82–87
4. Bhalotia G, Hulgeri A, Nakhe C, Chakrabarti S, Sudarshan S (2002) Keyword searching and browsing in databases using BANKS. In: *Proceedings of the 18th international conference on data engineering*. San Jose, CA, USA, February 26–March 1, 2002
5. Qin L, Yu JX, Chang L (2009) Keyword search in databases: the power of rdbms. In: *Proceedings of the 2009 ACM SIGMOD international conference on management of data*, pp 681–694
6. Hristidis V, Papakonstantinou Y (2002) DISCOVER: keyword search in relational databases. In: *International conference on very large data bases*, pp 670–681

7. Agrawal S, Chaudhuri S, Das G (2002) DBXplorer: a system for keyword-based search over relational databases. In: ICDE, pp 5–16
8. Liu F, Yu C, Meng W, Chowdhury A (2006) Effective keyword search in relational databases. In: Symposium on principles of database systems conference, pp 563–574
9. Zhang C, Naughton J, DeWitt D et al. (2001) On Supporting containment queries in relational database management systems. In: Symposium on principles of database systems conference, pp 425–436
10. Shaoul C, Westbury C (2011) A USENET corpus (2005–2010), Edmonton, AB: University of Alberta. <http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html>

Real-Time Subscriber Session Management on 5G NSA Wireless Network Systems



Kwan Young Park and Onur Soyer

Abstract 5G mobile network systems are known for their speed, security and durability. Speed and durability are rely on to security. To be able to realize these three core points, performing session managing in real-time is very important. In this paper we look in depth and discuss to achieve high performance session management on 5G NSA mobile networks.

Keywords 5G wireless networks · Security · Denial-of-Service · Session management · Distributed systems

1 Introduction

The vision of 5G wireless networks lies in providing very high data rates and higher coverage through dense base station deployment with increased capacity, significantly better Quality of Service (QoS), and extremely low latency [1]. In terms of QoS, system must be durable, reachable and secure all the time.

5G NAS wireless networks is introduced to public use in 2019 in Korea. Starting from that day number of 5G NSA subscribers reached 1 million in 69 days [2]. It is clear that the number of subscribers of 5G service will reach significant amount very soon. Subscribers of 5G wireless networks are users who are given access to network with a sim card. Sometimes sim cards might be embedded into the device itself. When a subscriber turn on the phone, it starts to communicate to the network for new session. Sessions in 5G NSA (Non-standalone) is the core point to handle subscriber connection and authorization to the network.

Each time a data sent from user's device, session status is the first to check before moving to the next step. Thus this is the first step to detect malicious attacks, such

K. Y. Park (✉) · O. Soyer
Mobigen Co., Ltd, C-16th Floor, 128, Beobwon-Ro, Songpa-Gu, Seoul 05854, Republic of Korea
e-mail: kypark@mobigen.com

O. Soyer
e-mail: onrsyr@gmail.com

as Signaling DoS [3], to the system. Mainly the aim of these attacks are to damage the system and to damage subscriber in terms of QoS and bill.

In the next sections, we will talk about how session manager works.

2 Distributed Session Management

System is composed of multi-server. Each server contains Kafka processes and for in-memory calculation each server has Redis processes. For the sake of real-time processing speed, all modules are developed in GO with multiprocessing.

Streaming GTP-U, GTP-C data are grouped from 0 to 9 and each group is kept in its own Kafka Topics. Producers are responsible to stream data from data pool and put it into Topics. After that each group of GTP data is consumed by Consumer processes and send to Session Management Modules. Each Consumer only consumes from its assigned topic which is the latest of number of IMSI (Figs. 1, 2 and 3).

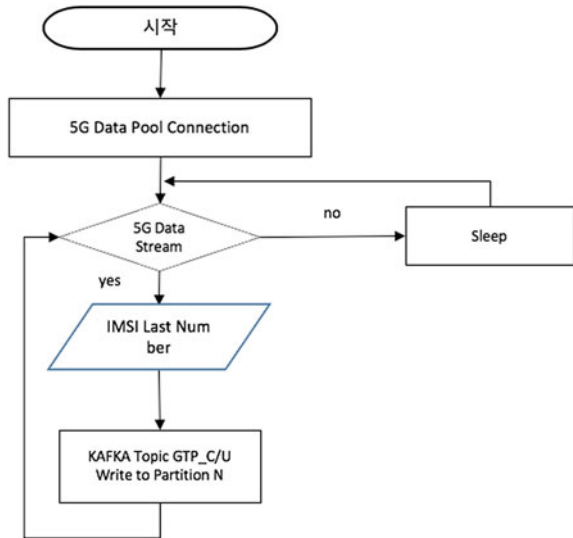
Consumer sets data to Redis in order and Session Handling Methods to get each data one by one. Data in Redis is kept as Key, Value.

IMSI, EBI and Request Type values are parsed from GTP data. At first, we create a unique key value using the following formula;

$$KEY = IMSI + EBI \tag{1}$$

After calculation we need to check if the key is already exists to determine Session status. If this is the first user attempt to connect to network, then key must not exist. If key cannot be found, it is added to Redis and GTP data as its value. In the next

Fig. 1 Distributed session manager handling flow



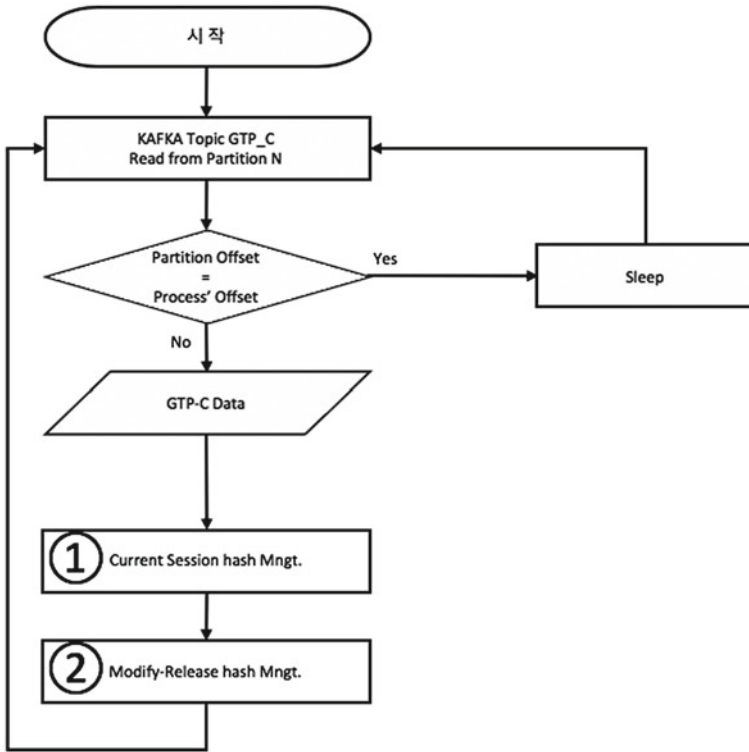


Fig. 2 Session traffic stream handling module

Key	Value								
IMSI + EBI	Timestamp	Message Type	APN	MME_TEID_C	SGW_TEID_C	SrNB_CU_TEID_U	SGW_TEID_U	SrNB_CU_ID	pdn_ipv4 Or pdn_ipv6

Fig. 3 Session map

step, Request Type control logic checks whether Request Type is ‘Delete Session’ type. If it is then the user’s session is deleted and logic ends Fig. 4.

When Session Type is ‘Modify Bearer’ or ‘Release Bearer’, session must be checked whether it exists. If not, then add the key with value. Each of Modify and Release Bearer requests are logged in a map. After each Release and Modify request is received, request count is increased by one and inserted to the map (Fig. 5).

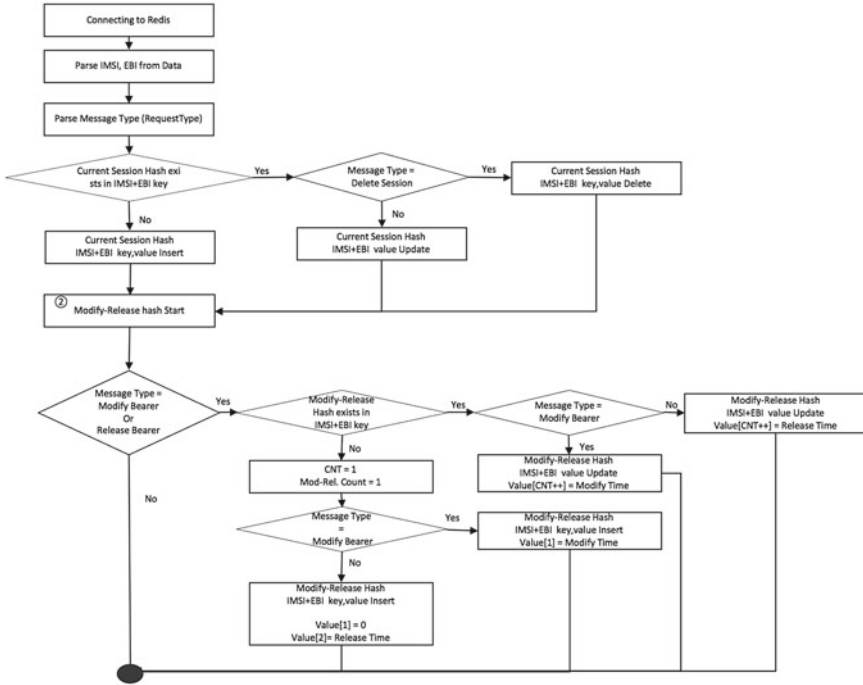


Fig. 4 Session handling module

Key	Value (array)							
IMSI + EBI	Mod-Rel Count	Modify Time 1	Release Time 1	Modify Time 2	Release Time 2	...	Modify Time N	Release Time N

Fig. 5 Modify/Release map

3 Results

We have used 3 nodes to perform our tests. Specs of each node are identical. 12 Core Intel(R) Xeon (R) CPU E5645 @ 2.40 GHz, 125 GB of RAM, 2 TB. of disk space. On each node, Kafka, Redis, Zookeeper and 5G Stream Session Handler are installed on all three nodes. Kafka Topic is set with a replication-factor 1.

After all processes are executed, data is streamed continuously for 20, 60 and 120 s. Size of streaming data is 300 byte. We took a record of each result also shown in the table below (Tables 1 and 2).

Table 1 Unoptimized random session data processing performance

Count/sec	20 s	60 s	120 s
Processed data count	9,352	9,983	9,660

Table 2 Avg. session data processing performance

Count/sec	20 s	60 s	120 s
Processed data count	31,187	32,172	31,810

According to the results, reading and writing task performance increase steadily up to 120 s which is its peak point. After that it continuous at the same rate. We observed that reading speed is around 1.2 million data per second and writing speed is 300 thousand data per second. However, after enabling session module, we observed that data processing rate drop to 30,000 data per sec which is expected due to heavy work.

For unoptimized random session data processing, the performance result is almost 3 times less than the optimized session data processing.

4 Future Works

This approach described in this paper belongs to an ongoing project. Our plan is to implement other types of attack on 5G networks system until the end of the project. Currently our main drawback is the lack of data. When we are able to get a stream of 5G data sample, we will perform more tests to make this approach more complete.

In the next step, we will implement Signaling DoS detection to prevent attacks. Also we will containerized each of the modules presented in this paper for the sake of modularity, easy management, versioning and security.

Acknowledgements This work was supported by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00793, Intelligent 5G Core Network Abnormal Attack Detection and Countermeasure Technology Development)

References

1. Agiwal M, Roy A, Saxena N (2016) Next generation 5G wireless networks: a comprehensive survey. *IEEE Commun Surv Tutor* 18(3):1617–1655, 3rd quarter
2. <http://www.businesskorea.co.kr/news/articleView.html?idxno=32809>, Accessed 16 Aug 2019
3. Jang et al. (2014) *J Adv Comput Netw* 2(3)

PCA and K-means Based Genome Analysis for *Hymenobacter* sp. PAMC26628



Ermal Elbasani, So-Ra Han, Tae-Jin Oh, Bongjae Kim, and Jeong-Dong Kim

Abstract The advancement of big data in biology, has made that the computational tools will become increasingly important and will be widely incorporated with various of analysis stream. Bioinformatics and particularly DNA sequence analysis is a challenging and trend topic. This study, use Principle Component Analysis and K-means cluster techniques are used to locate the genes within a genome sequence of the bacteria *Hymenobacter* sp. PAMC26628. The method is able to discover pattern in sequence and identify gene position in the genome, without prior known homology or gene annotation. Additionally, when the 64-dimensional space of codon probability distribution is applied for the first two and three principle components, a seven-cluster structure has resulted.

Keywords PCA · K-means · Genomic analysis · *Hymenobacter* PAMC26628

E. Elbasani · B. Kim · J.-D. Kim (✉)

Department of Computer Science and Engineering, Sun Moon University, Asan, Korea
e-mail: kjdvhu@gmail.com

E. Elbasani

e-mail: ermal.elbasani@gmail.com

B. Kim

e-mail: bjkim0422@gmail.com

S.-R. Han · T.-J. Oh

Department of BT-Convergent Pharmaceutical Engineering, Sun Moon University, Asan, Korea
e-mail: 553sora@hanmail.net

T.-J. Oh · J.-D. Kim

Genome-Based BioIT Convergence Institute, Asan, Korea

T.-J. Oh

Department of Pharmaceutical Engineering and Biotechnology, Sun Moon University, Asan, Korea

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_13

1 Introduction and Background

The emergence of modern medicine has fundamentally changed the nature of human existence. Thanks to the continuous development of medicine, the life expectancy of patients in developed countries has been greatly increased. Many diseases, among them hepatitis C, that were considered largely untreatable 10 years ago, are now treatable. The development of genetics can cure many diseases, and it can be treated immediately. With the innovation of diagnosis and measurement, doctors can specifically identify and target human diseases [2].

Fields that rely heavily on a single human observation now employ data sets that cannot be analyzed manually. Machine learning is now used daily to classify images of cells. The results of this high-level computational model are used to identify and classify tumor cases and evaluate the effectiveness of treatments for underlying diseases.

1.1 Genome Analysis

In the 1980s biologists would conduct single experiments and produce single results. This type of data can be manipulated manually with the support of simple calculators. Looking in prospective in current biology, it could generate millions of experimental data points in a day or two. Experiments such as gene sequencing, which can generate huge data sets, are not much expensive and easy to possess.

With the development of gene sequences, a database has been established linking individual genetic codes to many health-related outcomes, including genetic diseases such as diabetes, cancer and cystic fibrosis. Scientists are using computational techniques to analyze and mine data to develop an understanding of the causes of these diseases and to use that understanding to develop new treatments [1, 3]. For coding regions detection in genome, several methods have been applied among the most used are based on sequence comparison and the so-called ab ignition method [9]. Modern DNA sequencing has brought forth the extensive amount of genomic data. Computational methods, such as FASTA, BLAST, Hidden Markov Models, Interpolated Markov Models and Information Theory [4], provide efficient and reliable means towards analyzing these data sets and gene prediction.

1.2 K-means Clustering via Principal Component Analysis

Principal component analysis (PCA) is a widely used statistical method designed to reduce unsupervised dimension. K-means clusters are data clusters used for undirected learning tasks. The results of size reduction go beyond previous explanations of noise reduction and provide new insights into the observed effects of PCA base

data reduction. Mapping data points into a higher dimensional space via kernels, we show that solution for Kernel K-means is given by Kernel PCA. The results of this report suggest an effective technique for K-means clustering [6].

2 Hymenobacter sp. PAMC26628 Genome Analysis with PCA and K-means

PCA and K-means methods are able to determine genes with unknown homology, without context related to bacteria family, whereas the BLAST and FASTA algorithms can identify new genes with similar homology to known genes, the use of phylogenetic tree [8], This study is focused in Statistical methods that are particular useful for prokaryotic genomes, which are typically compact (10–20% noncoding DNA).

This study tests the Hymenobacter sp. PAMC26628 genome sequence form NCBI data bank. PCA is performed on the DNA sequence and fragment length $N = 300$, but varying codon length $n \in 1, 2, 3, 4$. Figure 1 clearly shows that DNA has an underlying structure and symmetry only in the $n = 3$ case evidently forming 7 clusters structure compare to figures when singlet, duplet and quadruplet codons. This is consistent with the genetic code, which describes how information is encoded in nonoverlapping triplets [5].

K-means is then used to identify seven distinct clusters in the flower-like structure. The correct phase-shift of the data is identified by looking at the mutual information of each fragment in the genome is shown Fig. 2.

The location of each of these data points in the DNA sequence is in Fig. 3. The location of each of these data points (e.g. each fragments) in the DNA sequence is in Fig. 3. There are assigned different colors for the cluster points. The cluster vary from the genes that correspond. Some gene (sequence fragments) overlap in a way that the information can be read with the correct shift, whereas other clusters contain the shifted information. The problem stays to find the cluster corresponding the correct shift. Connecting the centers of each of the clusters reveals an orbit with approximate C3 symmetry. These orbits have been identified in [7] as corresponding to the 3 different phase shifts in the forward and backward strands of the DNA. The mutual information (1) is used to calculate the correct phase shift, as well as determine the information content in the central cluster. The mutual information measures the information in each fragment

$$M = \sum_{ijk} f_{ijk} \log \left[\frac{f_{ijk}}{p_i p_j p_k} \right] \quad (1)$$

where p_k for $k \in \{A, C, G, T\}$ corresponds to the probability of observing the k -nucleotide and f_{ijk} is the probability of observing the codon triplet ijk for $i, j, k \in \{A, C, G, T\}$. When one or more nucleotides uniquely determines a codon, mutual

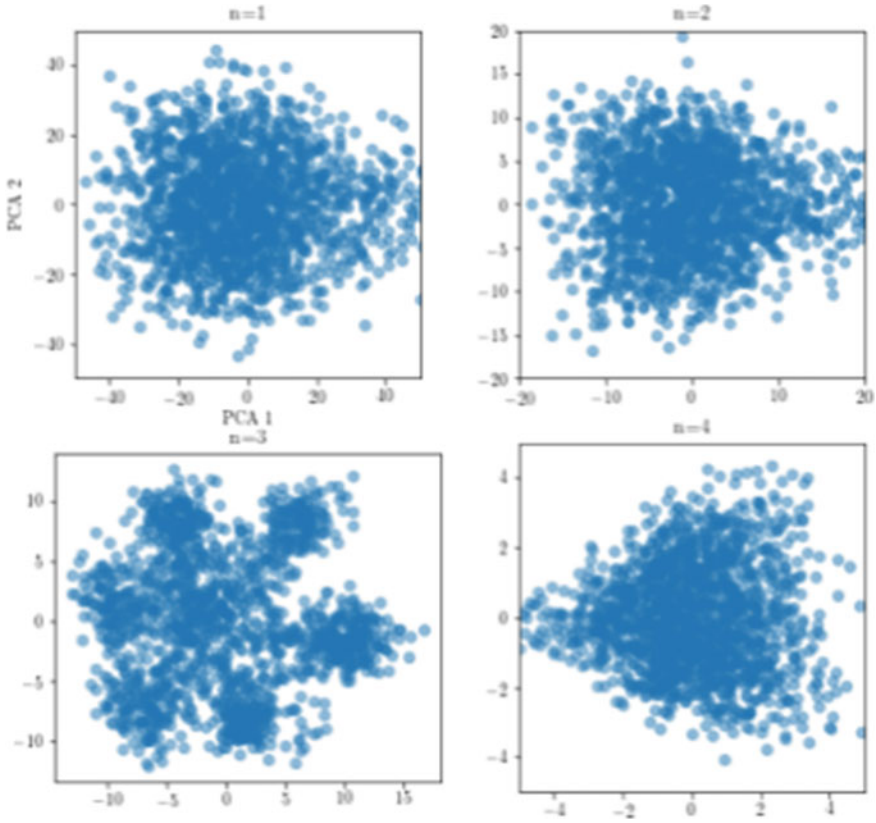


Fig. 1 Projected data along first two principle components for varying codon length of $n = \{ 1, 2, 3, 4 \}$. Only the overlapping triplet’s case ($n = 3$) has an underlying structure and symmetry

information is maximized. If p_i is a random variable that is independent, each codon Probability is simply the multiplication of each nucleotide probability, $f_{ijk} = p_i p_j p_k$. At this limit, all information regarding the nucleotide location is lost and $M = 0$. The non-coding areas of the bacterial genome are well represented by the state of maximum entropy.

3 Result and Discussion

On the basic Biology knowledge, from the experiment above, in order to verify more accurately, the mutual information was measured for a randomly generated. Figure 4 shows the mean mutual information per cluster. Cluster 1 has the minimum mutual information, which suggests the data in this region corresponds to fragments in the noncoding regions of the DNA sequence. In Fig. 5 shows that clusters 6 and 7 have the

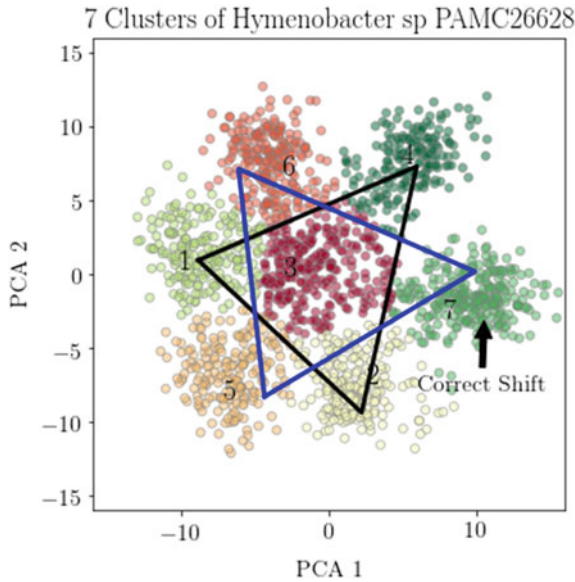


Fig. 2 K-means separates the flower-like structure into seven distinct clusters



Fig. 3 Cluster number of each fragment within the DNA sequence

highest amount of mutual information. To verify which cluster is the correct phase shift, the average number of stop codons per cluster is measured. So, the correct triplet probability, will have the minimum number of stop codons TAA, TAG and TGA, in a gene stop codons can appear only once to terminate the transcription.

It is clear from this figure that cluster 7 is the correct shift of the forward strand, as it has the minimum number of total stop codons. Therefore, it is highly likely

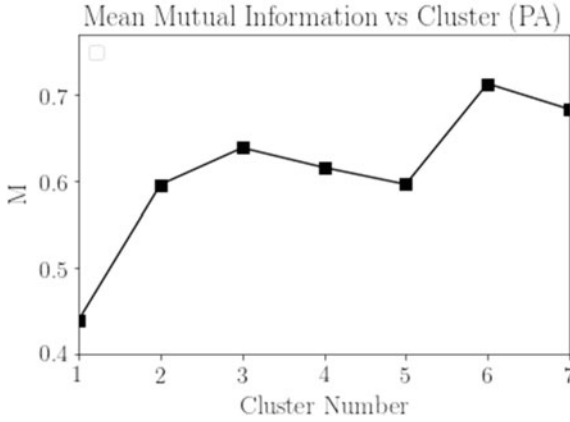


Fig. 4 The average mutual information per cluster. The figure suggests that the correct phase shift is either cluster 6 or 7, and the non-coding region is likely cluster 1

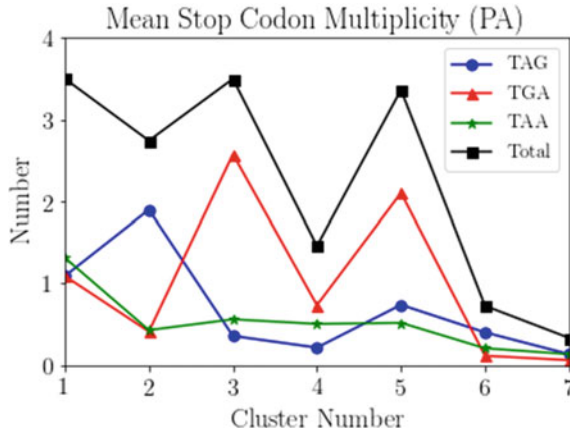


Fig. 5 The average number of stop codons per cluster

the PCA provides a meaningful representation of the data and that the flower-like structure correlates with the information embedded within the DNA sequence.

From the result above we are able to receive information from a simple clustering method in DNA analysis. This demonstrates that from analyzing only one genome sequence can detect and categories which fragments of sequence with high probability will find functionalities and easily to be annotated for later research and application. Distinguishing similarities in the DNA code using simple methods in difference with other current methods use based on homogeneous relation of the sequence.

4 Conclusion

This study provides an insight of using PCA and K-means as tool for analyzing genome analysis. These methods still have advantages to compared complex classification methods when come to implementation and prior data required to get knowledge from data.

Some limitation is to be mentioned when use PCA with K-means, despite the good clustering result, it is needed to specify K every time the algorithm run and is sensitive to outliers. In addition, this work was able to perform experiment based on bacteria genome sequence *Hymenobacter* sp. PAMC26628. On this study we were able to receive result by using cluster methods like PCA with K-means, that bring us to see also on the unsupervised learning complex methods. As a future work is to advance the research on neural networks to deal with network data, which is a common data type in bioinformatics. generative networks, Generative adversarial networks (GAN) and Variational autoencoder (VAE), which can be useful for biological and protein or drug design.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1058394), and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the National Program for Excellence in SW) (2018-0-01865) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Kumar M, Arora S, Pal A, Johri P (2016) A survey of big data analytics in healthcare. *INROADS Int J Jaipur Natl Univ* 5(1s):239
2. Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12(10):931–934
3. Alipanahi B, Delong A, Weirauch MT, Frey B (2015) Predicting the sequence specificities of DNA and RNA-binding proteins by deep learning. *Nat Biotechnol*
4. Leung MK, Delong A, Alipanahi B, Frey BJ (2016) Machine learning in genomic medicine: a review of computational problems and data sets. *Proc IEEE* 104(1):176–197
5. Krutovsky KV (2016) Seven-cluster structure of larch chloroplast genome. *J Siberian Federal Univ Biol* 8(3):268–277
6. Afrin F, Tabassum M (2018) Comparative performance of using PCA with K-means and fuzzy C means clustering for customer segmentation
7. Pecora LM, Sorrentino F, Hagerstrom AM, Murphy TE, Roy R (2014) Cluster synchronization and isolated desynchronization in complex networks with symmetries. *Nat Commun* 5:4079
8. King KM, Van DK (2018) Building (viral) phylogenetic trees using a maximum likelihood approach. *Current Protocol Microbiol* 51(1):e63
9. Picardi E, Pesole G (2010) Computational methods for ab initio and comparative gene finding. *Methods Mol Biol* 609:269–284

On Invariance of Concept Stability for Attribute Reduction in Concept Lattice



Fei Hao, Erhe Yang, Lantian Guo, Aziz Nasridinov, and Doo-Soon Park

Abstract Formal Concept Analysis (FCA) methodology, as an efficient knowledge representation and knowledge discovery tool and has been widely used in various fields, such as data mining, expert systems, and others. Knowledge reduction is an essential issue for knowledge discovery. This paper focuses on attribute reduction in FCA and explores the internal relation between concept stability and attribute reduction. By observing the concept stability of concepts in original concept lattice and reduced concept lattice, a theorem about the invariance of concept stability for attribute reduction in concept lattice is presented and proved mathematically. It is believed that the proposed theorem provides a novel solution for quick attribute reduction and benefit for other social system applications.

Keywords FCA · Attribute reduction · Concept stability · Invariance

This research was supported by the National Natural Science Foundation of China (Grant No. 61702317), MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00720) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and the National Research Foundation of Korea (No. 2017R1A2B1008421) and was also supported by the Natural Science Basic Research Plan in Shaanxi Province of China (2019JM-379).

F. Hao · E. Yang
School of Computer Science, Shaanxi Normal University, Xi'an, China
e-mail: feehao@gmail.com

L. Guo
School of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, China
e-mail: guolt0211@gmail.com

A. Nasridinov
Department of Computer Science, Chungbuk National University, Cheongju, Korea
e-mail: aziz@chungbuk.ac.kr

D.-S. Park (✉)
Department of Computer Software Engineering, Soonchunhyang University, Asan, Korea
e-mail: parkds@sch.ac.kr

1 Introduction

Formal Concept Analysis (FCA), as an effective methodology for knowledge discovery, has attracted much attention from the artificial intelligence community. At present, FCA has been widely used in machine learning [1], pattern recognition, expert systems [2], decision-making and data mining, and other related fields [3].

Knowledge reduction is an essential direction for knowledge discovery. In particular, attribute reduction in concept lattice provides powerful data reduction solution for knowledge discovery and data mining. Besides, attribute reduction as a key issue of FCA, its goal is to find minimal attribute set to maintain the invariant property. Generally, attribute reduction in FCA can be categorized as follows in terms of different reduction purpose: (1) **concept lattice complexity based attribute reduction**: this reduction aims to reduce the number of nodes in concept lattice and to delete the redundant concepts [4]; (2) **rule-based attribute reduction**: this reduction aims to maintain the rule sets unchanged [5]; (3) **lattice structure based attribute reduction**: this type of reduction is commonly studied in literature. It targets to keep the same structure of concept lattice and deletes the redundant attributes [6]. The attribute reduction in this work falls into the category (3).

This paper focuses on investigating the relation between concept stability in the original lattice and reduced lattice. For this, we introduce the definition of concept stability, which is used to measure the strength of dependency between intent and objects of extent. The main contributions of this paper are two-fold: (1) we present a theorem to prove that the concept stability is not changed during the attribute reduction of concept lattice (2).

The rest of this paper is organized as follows. Section 2 provides the preliminaries about Formal Concept Analysis methodology and elaborates the attributes reduction in concept lattice. By observing the concept stability of concepts in original concept lattice and reduced concept lattice, a theorem about the invariance of concept stability for attribute reduction in concept lattice is presented and proved mathematically in Sect. 3. Finally, Sect. 4 concludes this work.

2 Attribute Reduction in Concept Lattice

Before the presentation of attribute reduction in concept lattice, we briefly provide the preliminaries of FCA. FCA is a powerful methodology for describing the binary relationships between object and attribute and has been applied to many areas. Formally, a formal context is formulated as $C = (O, A, I)$ where O indicates the object set and A denotes the attribute set respectively, and the relation $I \subseteq O \times A$ refers to a binary relation between object and attribute. Generally, $o \in O$ and $a \in A$, $(o, a) \in I$ is interpreted as objective o has the attribute a .

For the sake clarity of formal concept lattice and its generated formal concepts, the following two operators are given [7].

(Operator for extracting the common attribute of objects subset X) For $X \subseteq O$, we define a set of common attributes of X,

$$X^\uparrow = \{ a \in A \mid (x,a) \in I, \forall x \in X \} \quad (1)$$

(Operator for extracting the common objects of attributes subset Y) For $Y \subseteq A$, we also define a set of common objects of Y,

$$Y^\downarrow = \{ o \in O \mid (o,y) \in I, \forall y \in Y \} \quad (2)$$

In a formal context $C = (O, A, I)$, for $X \subseteq O, Y \subseteq A$, if $X^{\uparrow\downarrow} = Y$, then this pair (X, Y) is called as a concept where X, Y are the extent and intent of the concept. Let $\text{Ext}(C)$ be the set of extents w.r.t. the formal context C. With the above operators, a concept lattice $L(C)$ can be defined as concepts organized according to a special hierarchical partial order, i.e., $(X_1, Y_1) \leq (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow Y_1 \supseteq Y_2)$.

Suppose that $C = (O, A, I)$ is a formal context, $D \subseteq A, I_D = I \cap (O \times D)$, then $C_D = (O, D, I_D)$ is a formal context as well and regarded as a sub-formal context of C.

Theorem 1 Suppose $C = (O, A, I)$ as a formal context, if $D \subseteq A$, then $\text{Ext}(C) \supseteq \text{Ext}(C_D)$ holds.

Theorem 1 demonstrates that the set of extents of sub-formal context is contained with the set of extents of original formal context.

Definition 1 Let $C_1 = (O, A_1, I_1)$ and $C_2 = (O, A_2, I_2)$ be two formal contexts, $L(C_1)$ and $L(C_2)$ be the corresponding concept lattices. For any concept $(X_2, Y_2) \in L(C_2)$, there exists $(X_1, Y_1) \in L(C_1)$ that satisfies $X_1 = X_2$, then we say that $L(C_1)$ is a refined version of $L(C_2)$, denoted as $L(C_1) \leq L(C_2)$.

Particularly, if $L(C_1) \leq L(C_2)$ and $L(C_2) \leq L(C_1)$ hold simultaneously, then the concept lattices of $C_1 = (O, A_1, I_1)$ and $C_2 = (O, A_2, I_2)$ are isomorphism, i.e., $L(C_1) \cong L(C_2)$.

Definition 2 Let $C = (O, A, I)$ be a formal context, $D \subseteq A$. If $L(C) \cong L(C_D)$, then D is the consistent set of C. Additionally, for any $d \in D, L(C_D) \not\cong L(C_{D-\{d\}})$, then D is the reduction of C.

For the sake of better presentation, this paper takes the above consistent set/reduction as the consistent/reduction of the concept lattice. As matter of fact, a consistent set D of a formal context C is a type of attributes set that maintains the invariant of extent set, i.e., $\text{Ext}(C) = \text{Ext}(C_D)$.

Example 1 Table 1 shows a formal context $C = (O, A, I)$, $O = \{1, 2, 3, 4\}$, $A = \{a, b, c, d, e\}$, the corresponding concept lattice and its reduction are presented as follows.

Table 1 A formal context $C = (O, A, I)$

	a	b	c	d	e
1	1	1	0	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	1	0	0

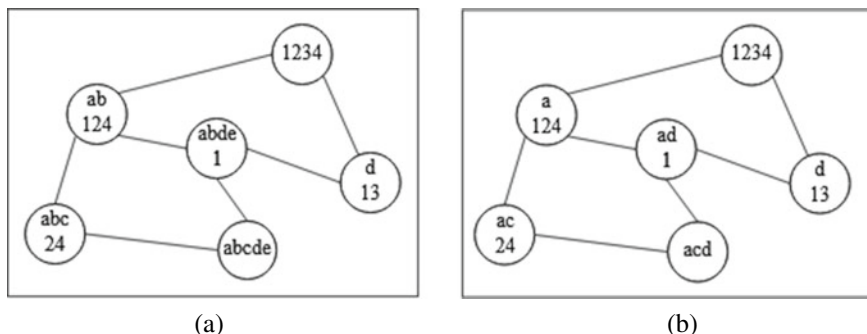


Fig. 1 Concept lattices $L(O, A, I)$ and $L(O, R_1, I_{R1})$

By invoking the concept lattice generation algorithm, the corresponding concept lattice for $C = (O, A, I)$ is shown in Fig. 1(a). Particularly, there are 6 concepts: $(1, abde)$, $(24, abc)$, $(13, d)$, $(124, ab)$, $(1234, \phi)$, $(\phi, abcde)$, denoted as $FC_i (i = 1, 2 \dots 6)$.

There are two reductions of this formal context, i.e., $R_1 = \{a, c, d\}$, $R_2 = \{b, c, d\}$. In this example, the attributes c and d are core attributes, attributes a and b are relatively necessary attributes, while attribute e is the redundant one. Figure 1(b) is the concept lattice of the formal context (O, R_1, I_{R1}) . Obviously, $L(O, R_1, I_{R1}) \cong L(O, A, I)$.

As can be seen from Fig. 1a, b, the extent of concepts are not changed, and the intent (i.e., attribute) are reduced from Fig. 1a. The structure of the lattices is an isomorphism.

3 On Invariance of Concept Stability for Attribute Reduction in Concept Lattice

In this section, we attempt to explore the relation between concept stability and attribute reduction in concept lattices. Firstly, we define concept stability. Further, the relation between concept stability and attribute reduction in concept lattices is figured out.

Definition 3 (*Stability Index*) [7] Let $c = (X, Y)$ be a concept of formal context $C = (O, A, I)$, then the intensional stability of (X, Y) is defined as follows.

$$\sigma(c) = \frac{|\{e \in \wp(X) | e' = Y\}|}{2^{|X|}} \tag{3}$$

In Eq. (3), intensional stability $\sigma(c)$ is used to measure the strength of dependency between the intent Y and the objects of the extent X . Specifically, it expresses the probability to maintain Y closed when a subset of noisy objects in X are deleted with equal probability. In other words, this index quantifies the amount of noise in the extent X and overfitting in the intent Y .

Inspired by this stability index, we think the redundant attributes will not affect the intensional stability $\sigma(c)$ when we delete some of them. Therefore, to answer our guess, it is necessary to investigate the relation between concept stability and attribute reduction in concept lattices.

Let us continue the Example 1, the stability for each concept in $L(O, A, I)$ can be obtained (as shown in Table 2) according to Definition 3.

Similarly, the stability for each concept in $L(O, R_1, I_{R_1})$ are also obtained as shown in Table 3.

By observing the above concept stability, we found that there is no any change from the original concept lattice to the reduced concept lattice which is obtained after deleting the redundant attributes. Therefore, the following theorem can be derived.

Theorem 2 Given a formal context $C = (O, A, I)$, its corresponding concept lattice is represented as $L(O, A, I)$. After deleting the redundant attributes, the reduced

Table 2 Concept stability of each concept in $L(O, A, I)$

Concept	Extent	Intent	Stability
FC_1	1	a,b,d,e	0.5
FC_2	2,4	a,b,c	0.75
FC_3	1,3	D	0.5
FC_4	1,2,4	a,b	0.375
FC_5	1,2,3,4	ϕ	0.0625
FC_6	ϕ	a,b,c,d,e	1

Table 3 Concept stability of each concept in $L(O, R_1, I_{R_1})$

Concept	Extent	Intent	Stability
FC_1	1	a, d	0.5
FC_2	2,4	a, c	0.75
FC_3	1,3	d	0.5
FC_4	1,2,4	a	0.375
FC_5	1,2,3,4	ϕ	0.0625
FC_6	ϕ	a,c,d	1

concept lattice is $L(O, R_1, I_{R_1})$. The stability of concepts is not changed during the reduction.

Proof Since the extracted concepts $(X, R), R \in R_1$ in reduced concept lattice $L(O, R_1, I_{R_1})$ have the same extent with original concept lattice $L(O, A, I)$. In addition, the redundant attributes are deleted from original attributes, that is to say, the $|\{e \in \wp(X) | e' = Y\}|$ equals to $|\{e \in \wp(X) | e' = R\}|$. Due to these, redundant attributes cannot affect the discernibility for a given extent X . Hence, we have the following equation:

$$\sigma(c) = \frac{|\{e \in \wp(X) | e' = R\}|}{2^{|X|}} = \frac{|\{e \in \wp(X) | e' = Y\}|}{2^{|X|}}$$

The above Theorem 1 holds.

4 Conclusions

This paper is the first work to explore the internal relation between concept stability and attribute reduction. First, this paper introduced the definition of stability index, then the concept stability of concepts in original concept lattice and reduced concept lattice are observed respectively; further, a theorem about the invariance of concept stability for attribute reduction in concept lattice is presented and proved mathematically. In the future, we will fully adopt the proposed theorem and devise a quick attribute reduction algorithm that can be used in other social system applications.

References

1. Shao M, Guo L, Wang C (2018) Connections between two-universe rough sets and formal concepts. *Int J Mach Learn Cybern* 9(11):1869–1877
2. Loia V, Orciuoli F, Pedrycz W (2018) Towards a granular computing approach based on formal concept analysis for discovering periodicities in data. *Knowl-Based Syst* 146:1–11
3. Hao F, Min G, Pei Z et al (2015) k-clique community detection in social networks based on formal concept analysis. *IEEE Syst J* 11(1):250–259
4. Wu W, Leung Y, Mi J (2008) Granular computing and knowledge reduction in formal contexts. *IEEE Trans Knowl Data Eng* 21(10):1461–1474
5. Li L, Mi J, Xie B (2014) Attribute reduction based on maximal rules in decision formal context. *Int J Comput Intell Syst* 7(6):1044–1053
6. Li M, Wang G (2016) Approximate concept construction with three-way decisions and attribute reduction in incomplete contexts. *Knowl-Based Syst* 91:165–178
7. Hao F, Sim DS, Park DS et al (2017) Similarity evaluation between graphs: a formal concept analysis approach. *J Inform Process Syst* 13(5):1158–1167
8. Ibrahim M, Missaoui R, Messaoudi A (2018) Detecting communities in social networks using concept interestingness. In: Proceedings of the 28th annual international conference on computer science and software engineering. IBM Corp, pp 81–90

A Study on Evidences Stored in Android Smartphones



Moses Kwak, Jisun Kim, Sungwon Lee, and Taenam Cho

Abstract Smartphones have become a necessity of modern people and have been used for various purposes beyond simple call and text transmission and reception. Therefore, various personal information has been stored in smart phones, which has become an important evidence of digital forensics. In this paper, we study the logs of messenger and web browser in Android smartphone which is widely used.

Keywords Digital forensics · Android · Smartphone · Security · Messenger · Web browser

1 Introduction

As smartphones are widely used as portable computers, various applications such as chatting, online shopping, and online banking are provided. Data maintained by these applications is not only sensitive user's personal information but also important evidence of digital forensics. Messengers are used as a major means of communication. These programs are used not only for simple chatting but also for web access for remittances and logins. Basic web browsers adopted by smartphones include Chrome, Firefox, Safari, Opera, and browsers developed by smartphone makers. In this paper, we investigate the information stored in the database used in the widely used web browsers and messengers and analyze the possibility of extraction.

M. Kwak · J. Kim · T. Cho (✉)
Woosuk University, Wanju, South Korea
e-mail: tncho@ws.ac.kr

M. Kwak
e-mail: rhkrahtp@naver.com

J. Kim
e-mail: rlawltjs122@gmail.com

S. Lee
Izerone Co. Ltd, Gunpo, South Korea
e-mail: jema10@daum.net

2 Experimental Environment and Object

2.1 Experimental Environment

To access a database stored on a smartphone, we clone the storage of smartphone and analyze the cloned image using forensics tools. We use specialized forensics tool for smartphone, AXIOM, and ADB (Android Debug Bridge) for direct access to databases as shown in Table 1.

2.2 Experiment Target

We used Samsung smartphone with Android as the experiment target. The detailed types are shown in Table 2.

We chose the five most widely used Web browsers and three Messengers for our experiments. The browsers and messengers and their version are shown in Table 3.

Table 1 Experiment Environment

Application	Version
AXIOM [1]	2.6
ADB	1.0.39

Table 2 Version of Target Device

Target	Version
Device	Galaxy A5 (2016)
Operating system	Android 7.0
Kernel	3.10.61–15139562
Knox	Knox 2.8
SQLite [2]	3.10.1

Table 3 Applications for experiment

Type	Application	Version
Web Browser	Chrome	76.0.3809.111
	FireFox	68.0.2
	Opera	53.0.2569.141117
	Safari	4.8.8
	Samsung browser	9.4.00.45
Messenger	Default SMS	4.1.28.68
	KakaoTalk	8.5.4
	Line	9.14.1

The sites accessed through Web browsers are two portal sites and three shopping malls.

3 Messenger

One of the services that are used frequently on smartphones is messenger. In particular, recent messengers support not only texts, pictures, files, but also online banking, and support link to web pages. In this paper, we investigated the text transfer logs of major messengers and web page connection through the text link.

3.1 Text Messages

When you use any messenger, if you stop the program while writing a message and go back to the app, you can see the messages you have entered. If so, these messages would be stored in areas other than “transmitted messages,” and this data could be evidence of digital forensics.

On default SMS, KakatoTalk [3] and Line [4], we had 3 experiments. (1) Sending a message (2) Activating another application (deactivating the messenger) after entering a message without sending it. (3) Terminating the messenger after entering a message without sending it. The data survey was conducted in parallel with AXIOM and ADB. The results of the experiments showed that not only the transmitted messages but also the untransmitted messages are stored in the database. However, the message is stored as a readable text or as an unreadable text as shown in Table 4.

3.2 Links to Websites

When the receiver access the website by clicking the received URL, the browsers used to access website for each messenger are shown in Table 5. In the case of KakaoTalk,

Table 4 Stored data in case text message transmission

Messenger	Action		
	Send message	Inactivate without message sending	Close without message sending
Default SMS	Readable	Unreadable	Unreadable
KakaoTalk	Unreadable	Readable	Readable
Line	Readable	Unreadable	Unreadable

Table 5 Stored web browser data

Messenger	Used web browser	Logs
Default SMS	Samsung browser	Stored
KakaoTalk	KakaoTalk browser	Not Found
Lie	Line browser	Stored

website access logs were not found. In the case of Line, related information that can infer URL was found.

4 Website Login Record

To get various useful services via website, we must put our credential information such as id and password on the login page. The database for access log of each browser is shown in Table 6. As in the messenger, we analyzed saved information before and after the user entered the login information and pressed the send button.

4.1 Experiment Results

We tried to log in to three portal sites (P1, P2 and P3) and two shopping sites (S1 and S2) and tested whether they record logs user's login information in databases. Like for the experiment in messenger, we observed logs for login ID and connection time before and after submitting login information. Unlike messenger, we couldn't find any information before submit. The password was not saved even when submitted. The stored sensitive information is shown in Table 7. As shown in Table 7, whether the data can be found depends on the target site or the browser used, and it is understood that the web page itself provides a function to prevent the storage or to prevent the web browser from storing.

Table 6 WebBrowser's dataBase

WebBrowser	DataBase
Chrome [5]	Web data
FireFox [6]	Formhistory
Opera [7]	Web data
Safari [8]	Not known
Samsung	Web data

Table 7 Stored login information for each browser

Browser	Target site	ID	First time	Last time	Count
Chrome/firefox	P1	Stored	Stored	Stored	Stored
	P2	Stored	Stored	Stored	Stored
	P3	Not found	Not found	Not found	Not found
	S1	Stored	Stored	Stored	Stored
	S2	Not found	Not found	Not found	Not found
Samsung/opera/safari	P1	Not found	Not found	Not found	Not found
	P2	Not found	Not found	Not found	Not found
	P3	Not found	Not found	Not found	Not found
	S1	Not found	Not found	Not found	Not found
	S2	Not found	Not found	Not found	Not found

5 Conclusion and Future Study

Most popular smartphone applications include messenger and web browser. Messengers allow users to access websites via links as well as send messages. Most websites require users to log in. We investigated the web browsers used to access websites via messengers. We also examined the login information stored in the smartphone before and after login for several popular websites. In addition, we examined the information stored before and after the message transmission through messengers. Experimental results show that information is stored before and after sending messages, depending on the messenger or website. The results of this experiment can help to collect digital forensic evidence through various channels. In the future, we will investigate foreign sites and study the relationship between web page design and browser and logged data.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03032637).

References

1. Magnet AXIOM, <https://www.magnetforensics.com/products/magnet-axiom/>
2. SQLite, <https://www.sqlite.org/index.html>
3. KakaoTalk, <https://www.kakaocorp.com/service/KakaoTalk?lang=ko>
4. Line, <https://line.me/ko/>
5. Chrome, <https://www.google.com/chrome/>
6. FireFox, <https://www.mozilla.or.kr/community/>
7. Opera, <https://www.opera.com/ko>
8. Safari, <https://www.apple.com/kr/safari/>

Divide the FCA Network Graph into the Various Community Based on the k -Clique Methods



Phonexay Vilakone, Min-Pyo Hong, and Doo-Soon Park

Abstract The current research trends many researchers apply the community detection method in the graph generated from the social network to create various real-world applications. The different result of used this method given the high performance and high accuracy, especially in the field of prediction the information which matches the information of the user requirement. Therefore, in this paper, we proposed a divide the FCA network graph into the various community based on the k -Clique method. We offer a brief description of the relevant techniques and present an innovative technology on how to create a community of users from the FCA network graph using the k -clique method.

Keywords Formal concept analysis · k -Clique · Network graph

1 Introduction

In recent years; research on the field of community detection in the social network or the large network graph has attracted more attention from the researchers. Many researchers try to develop on the different new technique on how to create the community from social network graph [1, 2] and apply the result of this method to the various real-world applications and the effect of using this method also given the high performance. However, there are a few research that tries to make the community from the network graph that created from the formal concept analysis and prove this method is more high performance [3]. Therefore; in this article, we proposed on a divide the

P. Vilakone · M.-P. Hong · D.-S. Park (✉)
Department of Computer Sciences and Engineering, Soonchunhyang University,
Soonchunhyang-Ro 22, Sinchang-Myeon, Asan-Si, Chungcheongnam-Do, South Korea
e-mail: parkds@sch.ac.kr

P. Vilakone
e-mail: xayus@yahoo.com

M.-P. Hong
e-mail: hmp4321@korea.kr

FCA network graph into the various community based on the k -clique methods. The main idea of this proposal is after we used users' personalized information to create the network graph based on the formal concept analysis method and its related software. Then, we used the k -clique method to detect the community from this graph. The result of the experimental give more beneficial, and researcher might use this method to their work.

The structure of this paper organized as follows. The next part describes the relevant to the work in Sect. 2. The idea of this paper will then explain in Sect. 3. Finally, we conclude our work and future study in Sect. 4.

2 Related Work

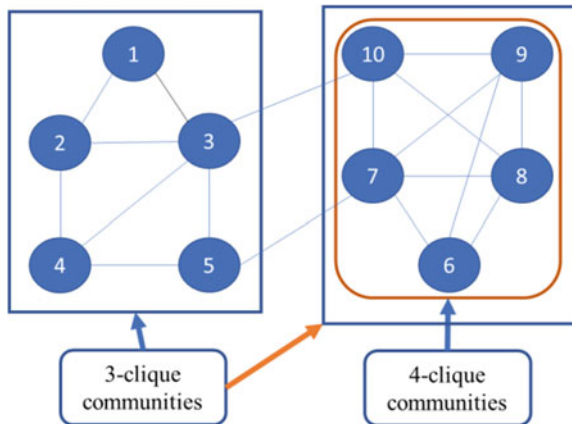
In this sector; the brief description of the relevant method that used in this paper will present, such as a formal concept analysis and k -clique method.

2.1 The k -Clique Method

The k -Clique method is a community detection in social network analysis and is a complete communication graph with k nodes. In general, the value of k is more significant than or equivalent to 3 [4].

In the Fig. 1, picture in the left-hand side; the set of nodes $\{1, 2, 3\}$, $\{2, 3, 4\}$, $\{3, 4, 5\}$, $\{6, 7, 8\}$, $\{7, 8, 9\}$, $\{8, 9, 10\}$, $\{7, 8, 10\}$ and $\{6, 8, 9\}$ represent groups of 3-clique because all nodes are completely connected to each other. Similarly, the picture in the right-hand side; the sets of nodes $\{6, 7, 8, 9\}$ and $\{7, 8, 9, 10\}$ represent groups of 4-clique.

Fig. 1 An example of a graph with its k -clique communities for $k = 4$ and $k = 3$



2.2 Formal Concept Analysis

Formal concept analysis (FCA) is a method for displaying knowledge, data management, and data analysis [5–9]. Formal concept analysis applied in many areas such as medicine, sociology, mathematics, psychology biology, or economics. The most exciting applications of FCA has expertise in computer science and can use for information retrieval, data mining, source code error correction, data analysis, and machine learning [10–12]. We expand on our previous paper [13]. The input data for the FCA method is displayed in a matrix format so that each row shows objects of the interest, and each column defined attributes. The matrix input element can assume only Boolean values, such as an object with or without specific properties. If an object has a unique feature, the “X” sign will be at the intersection of that row of the object and the column of that attribute. Otherwise, if the object does not have a specific property, the intersection of that object’s row and the column of that attribute will be empty.

3 Proposed Method

The process of gathering data and work processes for the guidance system is present in Fig. 2.

Process 1; In this process, the Movie lense Dataset [14] used as a training dataset. The 800 users’ ID assigned as an object and 30 different user’s personalized information assigned as an attribute. As stated earlier, an “X” marks in the formal context matrix on the intersection of the object and attribute, as shown in Table 1.

Fig. 2 The proposed method’s workflow

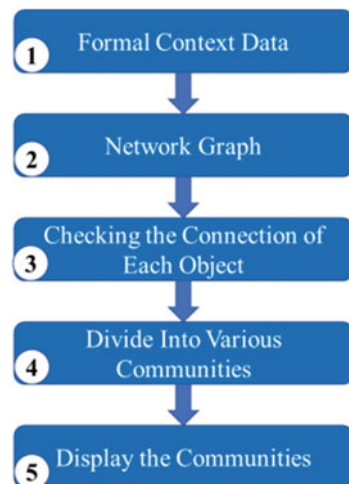


Table 1 Formal context of a Fundamentals of user’s personalized characteristic

	a	b	c	d	e	f	g	h	i	j	...	dd
1				x				x		x		x
2	x						x					x
3		x						x				
4			x			x				x		
5	x				x							x
...			x					x		x		
800		x				x						

In Table 1, at first row of a table, letters “a” refers to the user has age below 20 years, letters “b” refers to the user has age between 21 and 30 years, letters “c” refers to the user has age between 31 and 40 years, letters “d” refers to the user has age between 41 and 50 years, letters “e” refers to the user has age between 51 and 60 years, letters “f” refers to the user has age between 91 and 70 years, letters “g” refers to the user has age over 71 years, letters “h” refers to the user gender is male, letters “i” refers to the user gender is female, letters “j” refers to the user has occupation as administrator, letters “k” refers to the user occupation as artist, letters “l” refers to the user has occupation as lawyer, until letters “dd” refers to the user occupation as writer. In the first column of a table, the number from 1 to 800 is represented to the user id.

After preparing the training dataset then import this dataset to the Lattice Miner 2.0 program.

Process 2; when the formal context is displayed in the software then using lattice function to generate the network graph. The result of this process is shown in Fig. 3.

In Fig. 3, letters “o” to “f” was present the attribute of the user, the number “713” and other numbers which display in the Fig. 3 are represented to the user id and line are represent to the connection of user id and user attribute.

Process 3; After network graph has appeared, then function association rule in the Lattice Minor 2.0 program is used for checking the connection of each object. After that, the result of this process will export to the XML file. Table 2 shows a list of some association rules with the best support (minimal support of at least 10%).

Process 4; The XML file will Import to the R program. Then, the network graph from XML will create with the help of graph function in R as shown in step 2 of Algorithm 1 below. After that, the *k*-clique algorithm will use to detect a user in the network graph and divide them into the various community as a start from step 3 to step 5 of Algorithm 1 below. While the *k*-clique algorithm is running. The community will find and create from the value of *k* = 3 until the value of *k* can not found. The algorithms that use for community detection based on *k*-clique is shown below.

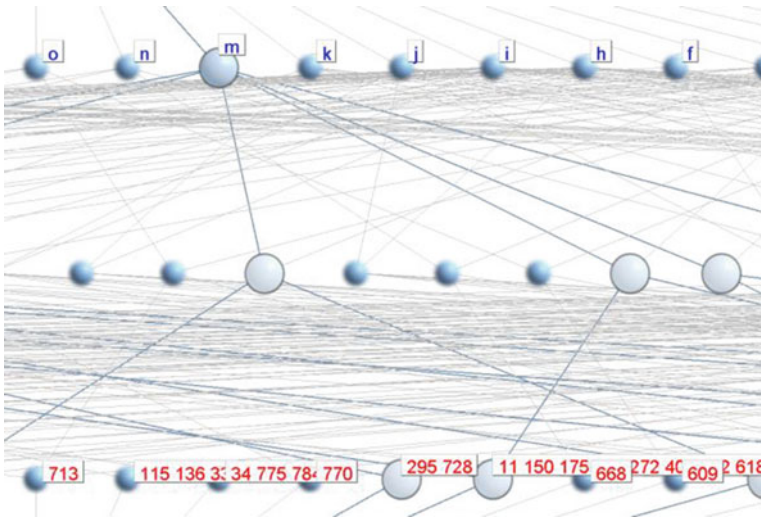


Fig. 3 Network graph

Table 2 List of association rules with minimal support of 10%

#	Antecedence	=>	Consequence	Support (%)	Confidence (%)
1	{h}	=>	{d, h, j, i}	12.24	19.31
2	{e}	=>	{j, k, l, bb}	15.72	23.31
3	{h, n}	=>	{b, j, k, m}	27.24	38.56
4	{c, h, n}	=>	{c, f, g, k}	18.07	27.15
5	{e, i, j}	=>	{h, u}	15.54	63.91
6	{c, h, m}	=>	{h, x, y}	17.32	69.57
7	{f, h, j}	=>	{b, g, k, l}	15.0	58.15
8	{b, dd, h}	=>	{h, k}	29.24	83.74
...	...	=>
N	{aa, b, h}	=>	{h, l}	19.47	52.59

Algorithm 1: Community Detection Based on *k*-Clique

Input: XML file

Output: Community

Step 1: my_network, community, k

Step 2: my_network <- graph.XML

Step 3: for k = 3 to n

Step 4: community <- clique(my_network, min = k, max = k)

Step 5: write.table(community, file = "community.txt")

Step 6: end

Table 3 Result of the community in case value of $k = 3$

Group id	User id in the group
1	user21, user63, user95
2	user97, user267, user95
3	user123, user156, user612
4	user591, user597, user791
5	user80, user221, user669
6	user51, user220, user749
7	user340, user288, user422
8	user07, user45, user723
9	user42, user375, user391
10	user561, user461, user727
...	...
1358	user703, user475, user603

Process 5; At the last operation, the result of communities has appeared. Some of the results from this process are shown in Table 3. In Table 3 shown the result in case of the value of $k = 3$. The total community when $k = 3$ is 1358 communities.

4 Conclusions and Future Study

In this paper, we presented a brief description, the significant concepts, and explained the main features of the formal concept analysis method. Besides that, a brief description of the k -clique process also present in this paper. In the end, we gave a short example showing how to create a network graph using the formal concept analysis. Then, we offer a brief explanation on how to divide a network graph which generates from formal concept analysis into the various communities. The advantage of the proposed method is to offer a new technique to the researcher on the community detection method in the FCA network graph, and we firmly believe that this idea will be beneficial to them. The goal of our future research work, we intend to prove the accuracy of the concept of using this method and for the quality of the efficiency in the field of recommendation system.

Acknowledgments This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00720) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and the National Research Foundation of Korea (No. 2017R1A2B1008421).

References

1. Hao F, Sim DS, Park DS, Seo HS (2017) Similarity evaluation between graphs: a formal concept analysis approach. *J Inf Process Syst* 13(5):1158–1167
2. Hao F, Park DS, Min G, Jeong YS, Park JH (2016) k -cliques mining in dynamic social networks based on triadic formal concept analysis. *Neurocomputing (Elsevier)* 209:57–66
3. Hao F, Park DS, Pei Z (2017) Detecting bases of maximal cliques in social networks. In: MUE2017, Seoul, Korea, pp 1–1
4. Hao F, Min G, Pei Z, Park DS, Yang LT (2015) K -clique communities detection in social networks based on formal concept analysis. *IEEE Syst J*. <https://doi.org/10.1109/jsyst.2015.3294>
5. Wille R (1982) Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival I (ed) *Ordered sets*. Reidel, Dordrecht-Boston, pp 445–470
6. Ganter B, Stumme G, Wille R (2005) *Formal concept analysis—foundations and applications*. Springer, Berlin, Heidelberg
7. Wolff KE. The first course in formal concept analysis. In: Faulbaum F (ed) *StatSoft '93*. Gustav Fischer Verlag, pp 429–438
8. Ganter B, Wille R (1999) *Formal concept analysis-mathematical foundations*. Springer, Berlin
9. Belohlavek R (2008) *Introduction to formal concept analysis*. Olomouc. Belohlavek.inf.upol.cz/vyuka/IntroFCA.pdf
10. Ganter B, Godin R (2005) In: *Third international conference on formal concept analysis, ICFCA 2005*. Springer
11. Kuznetsov SO, Schmidt S (2007) In: *Fifth international conference on formal concept analysis, ICFCA 2007*. Springer
12. Ferré S, Rudolph S (2009) In: *Seventh international conference on formal concept analysis, ICFCA 2009*. Springer
13. Vilakone Ph, Xingchang Kh, Joo WS, Park DS (2019) A network graph using the FCA method from user's personalized characteristic. In: *BIC2019*
14. Harper FM, Joseph AK (2015) The MovieLens datasets: history and context. *ACM Trans Interact Intell Syst (TiIS)* 5(4):19, Article 19

An Efficient Micro-Service Placement Scheme Based on Fuzzy System for Edge-Enabled Digital Signage Service



A.-Young Son, Yeon Soo Lim, and Eui-Nam Huh

Abstract With the rapid increase in the development of the Internet of Things, a large amount of data are expected to be generated, which result to in increased latency. To reduce the latency, service placement method has been researched for resource management from mobile devices to nearby edge server. However, most of the related studies did not provide sufficient service efficiency for multi-objective such as energy efficiency, resource efficiency, performance improvement. Also, most of the existing approaches did not consider various metrics. Thus, maximize performance and reduce cost, we consider multi-metric by combining decision method according to user requirements. In order to satisfy the user's requirement based on service, we propose an efficient service placement scheme based on Fuzzy-AHP, Finally, we prove the performance of the proposed scheme by using different placement schemes.

Keywords Cloud computing · Fuzzy system · Internet of thing · AHP · Service placement

1 Introduction

As the number of cloud providers and services increase, the composition of many services from different providers becomes more sophisticated in a real cloud environment. With the rapid increase in the development of Internet of Things (IoT) service, A large amount of data and micro-services such as digital signage are expected to be generated, which result to in increased latency. The digital signage market is expected

A.-Y. Son (✉) · Y. S. Lim · E.-N. Huh
Department of Science and Engineering, Kyung Hee University, Yongin, Republic of Korea
e-mail: ayths28@khu.ac.kr

Y. S. Lim
e-mail: 2014104139@khu.ac.kr

E.-N. Huh
e-mail: johnhuh@khu.ac.kr

to witness a CAGR of 8.0% by 2025, owing to increasing demand as per a study by Grand View Research, Inc [1]. Under these environments, by placing resource-rich nodes in close proximity to mobile or IoT devices, edge-enabled distributed computing is required more responsive services, along with higher scalability and lower latency than traditional cloud platforms [2, 3]. To address challenges, we propose an efficient service placement scheme based on Fuzzy in edge-enabled digital signage service.

Organization of the paper is as follows: In Sect. 2 analysis related studies to solve the limitation. In Sect. 3, we present proposed placement scheme. In Sect. 4, we proved performance with exiting scheme. The last section conclude this paper.

2 Related Work

The optimization techniques are one of the most popular methods for the resources management. By describing the predicted workload and the available resources as the constraints of the optimization problem, the optimal or the near optimal configurations of resources are found. Thus, the time complexity of the optimization techniques should be reasonable to find the optimal configurations according to the workload dynamics. As shown Table 1, previous work designed based machine learning techniques [3–5]. In order to solve the limitation, we will apply fuzzy system. We proposed micro-service placement scheme based on Multi-Criteria Decision-Making (MCDM) [6]. This method useful when the decision is applied in practice. Each metric affects the target machine selection for resource management. So we used this method in order to the selection of VM through the assigned priority based on MCDM. It used in many areas due to the ability to decision of alternative according to weight.

Table 1 Related work

Techniques	Advantage	Disadvantage
ARIMA	Simple	Dynamic workload
Bayesian theory	Simple	Inability to adapt to workload changes
Reinforcement	Dynamic workload	Performance Poor scalability in the large state space
Fuzzy	Modeling uncertainties and ambiguities	Complexity of rule set

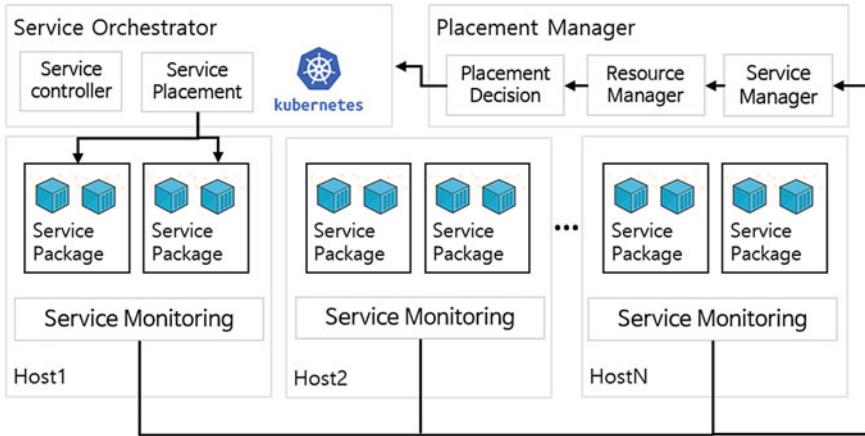


Fig. 1 Fuzzy-AHP system

3 Proposed Scheme

As shown in Fig. 1, it consists of two main modules: placement manager and service orchestrator

Placement manager is decide service placement based on information for resource status and service status. It consists of service manager, resource manager and placement decision.

- Service manager: in this module, we categorize service according to service type based requirements. After service categorization, it send to placement decision for categorization information.
- Resource manager: we analysis resource status based on resource monitoring
- Placement decision is designed based on fuzzy for making the decision on service placement situation whether or not. Depending on inputs such as CPU computation, RAM computation, it can control service placement efficiency.

In service orchestrator, we can action of placement based on result of placement manager considering of requirements. As shown Fig. 2, step of fuzzy-AHP has three processes.

4 Evaluation

In this section, we show the result of evaluation for proposed scheme and existing algorithm. To compare efficiency, we use simulation tool called Cloud sim 3.0. Table 2 shows evaluation scheme explanation. We consider three workloads defined in terms

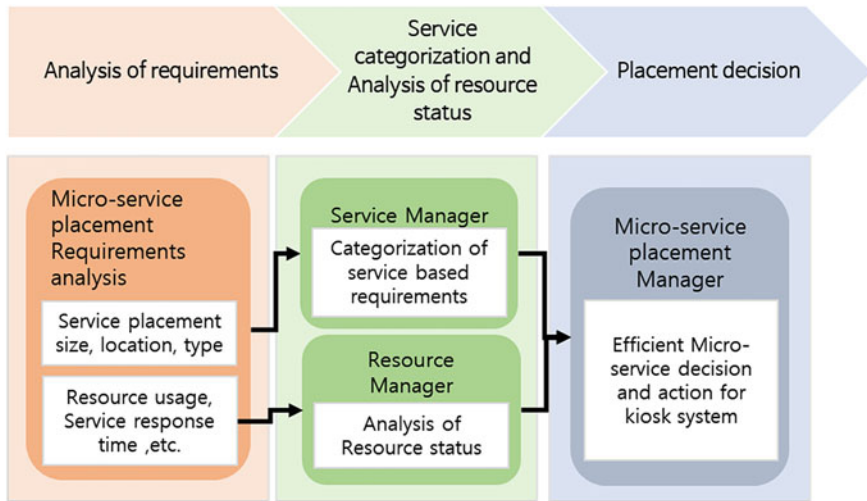


Fig. 2 Step of Fuzzy-AHP system

Table 2 Existing scheme

Scheme	Description
ARIMA	ARIMA stands for Autoregressive Integrated Moving Average models. ARIMA is a forecasting technique that future values of a series based entirely on its own metric
LSTM	LSTM is type of learning algorithm based on time series data

of the number of concurrent users accessing the application: low usage (100 users), normal usage (300 users) and heavy usage (600 users).

We selected target machine based on Fig. 2 according to the result of Fuzzy-AHP score. According to case, We also compared the results of the execution time and for proposed scheme with ARIMA and LSTM. In Figs. 3, 4 and 5, we can see the decrease in execution time less than LSTM. Thus, our scheme is proved efficient service placement more than existing scheme.

5 Conclusion

This research aims to make a scheme based decision system on distributed cloud computing environment. Users move frequently, thus it is necessary for data centers to periodically service placement to improve the resource efficiency of the system. According to evaluation result, our proposed scheme is able to efficiently manage the resource.

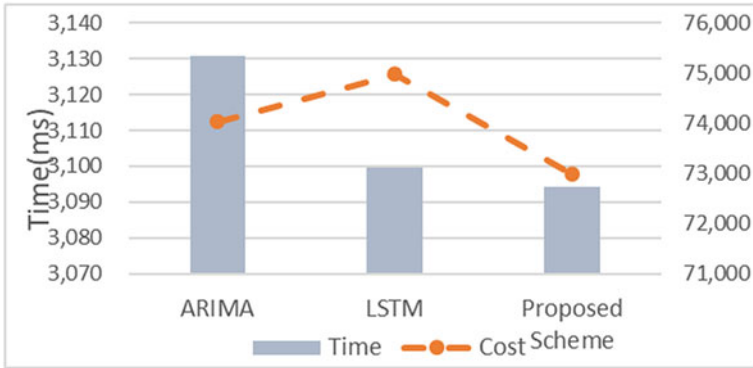


Fig. 3 Case1 (low): time and cost

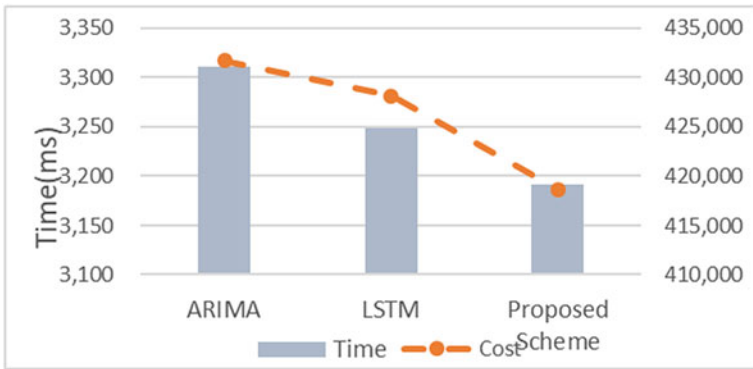


Fig. 4 Case2 (normal): time and cost

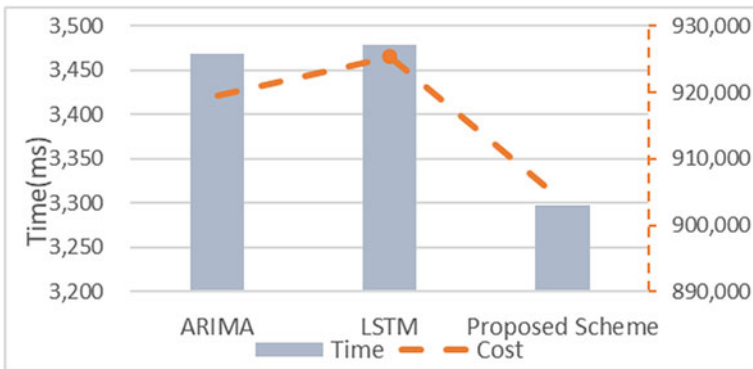


Fig. 5 Case3 (heavy): time and cost

Acknowledgements This work was supported by Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01615, Developed digital signage solution for cloud-based unmanned shop management that provides online video advertising) and the MIST (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2017-0-00093), supervised by the IITP (Institute for Information and communications Technology Planning and Evaluation) Corresponding author: Eui-Nam Huh.

References

1. Wang S, Xu J, Zhang N, Liu Y (2018) A survey on service migration in mobile edge computing. *IEEE Access* 6:23511–23528
2. Chen M, Li W, Fortino G, Hao Y, Hu L, Humar I (2019) A dynamic service migration mechanism in edge cognitive computing. *ACM Trans Internet Technol (TOIT)* 19:30
3. Machen A, Wang S, Leung KK, Ko BJ, Salonidis T (2017) Live service migration in mobile edge clouds. *IEEE Wirel Commun* 25(1):140–147
4. Abdah H, Barraca JP, Aguiar RL (2019) QoS-Aware service continuity in the virtualized edge. *IEEE Access* 7:51570–51588
5. Kikuchi J, Wu C, Ji Y, Murase T (2017) Mobile edge computing based VM migration for QoS improvement. In: 2017 IEEE 6th global conference on consumer electronics (GCCE), pp 1–5. IEEE
6. Machen A, Wang S, Leung KK, Ko BJ, Salonidis T (2017) Live service migration in mobile edge clouds. *IEEE Wirel Commun* 25:140–147

Rethinking Blockchain and Decentralized Learning: Position Paper



Sandi Rahmadika and Kyung-Hyune Rhee

Abstract Blockchain technology and decentralized learning are attracting growing attention. Most existing methods of machine learning is in the centralized form which relies upon the third party in terms of the raw datasets and mining resources. Blockchain solves world centralization problems that keep the system secure through complex mathematical computations puzzle solved by blockchain miners. Concurrently, decentralized learning such as federated model allows the user to collaboratively access the updated prediction model without revealing the training data to the public. By doing so, it provides less power consumption, lower latency that respects the user's privacy concern. Therefore, we study the extent to which these two technologies can be applied in the real world for faster convergence without compromising user's security.

Keywords Blockchain technology · Decentralized learning · Federated model · User's privacy

1 Introduction

In recent years, blockchain technology has become front and center of the latest influence technology that affects transactions systems in the real world. It is widely discussed with various intuitive applications which are being managed by blockchain network architecture. Unlike the centralized system in general, the transactions on the blockchain network are executed by the system without the third party involvement. The emergence of this technology resolves the problem of centralized systems

S. Rahmadika
Interdisciplinary Program of Information Security, Graduate School, Pukyong National University, Busan, South Korea
e-mail: sandika@pukyong.ac.kr

K.-H. Rhee (✉)
Department of IT Convergence and Application Engineering, Pukyong National University, Busan, South Korea
e-mail: khrhee@pknu.ac.kr

in terms of reliance on third parties to manage every transaction that occurs [1]. Furthermore, blockchain is secure by design since it is applying cryptographic protocols which are embedded with an identical timestamp for each transaction. Explicitly speaking, blockchain is tamper-proof from the attacker's attempt in modifying the history of the transaction [2].

Along with the popularity of the decentralized system pioneered by the blockchain, decentralized learning in the artificial intelligence (AI) area has also been developed lately. Decentralized learning system provides a breakthrough in the machine learning area since it changes the centralized form of raw data into a decentralized system [3] such as federated learning (FL) model which is introduced by Google AI research team in 2016 as a solution to improve the efficiency [4]. The emergence of decentralized learning can be interpreted as an intersection of edge computing, internet-of-things (IoT), and on-device AI environment.

According to the statistical data of IoT analytic [5], the IoT devices will continue to grow until it reaches 10 billion in 2010 and 22 billion in 2025. Due to the number of connected devices is growing all the time in the IoT environment, decentralized learning is not impossible to be realized on a large scale to overcome the problems in a centralized system. Therefore, in this paper, we explore the extent to which decentralized learning can be applied to overcome problems in a decentralized system. Collaborative models of decentralized learning and blockchain are also elaborated, along with other essential matters.

The structure of the paper is as follows. In Sect. 2, we present a decentralization in Blockchain technology. We select a number of points that we consider essential to be presented in this section. The emergence of decentralized learning whipped up the motivation for researchers to develop this technology. One way to develop it is to utilize blockchain technology. These are discussed in Sect. 3 and Sect. 4, respectively. Finally, Sect. 5 concludes the paper.

2 Decentralization in Blockchain

Blockchain peer-to-peer (P2P) network appears to the public is an architecture network that manages workload partitions between the parties in the system. It keeps a tamper-proof transaction record. The parties are also called as peers or nodes which are connected in the same network. The nodes manage every transaction that occurs in the blockchain network [6].

The consensus mechanism is essential in the blockchain network. The consensus is used by miners to validate every transaction that appears. Table 1 presents an overview of blockchain consensus after the emergence of proof-of-work, which is first adopted by Bitcoin.

Intuitively, blockchain is a chain-shaped data structured known as a chain of blocks [8]. It allows data being distributed to all nodes in the same network. When the node receives a new notification of the transaction in the network, the node verifies the incoming message as follows:

Table 1 The blockchain consensus after the emergence of proof-of-work [7]

	Permission-less access	Permissioned access
Decentralized validation	<ul style="list-style-type: none"> • Proof-of-work • Proof-of-stake • Proof-of-work based derivatives • Federated Byzantines agreement 	<ul style="list-style-type: none"> • Proof-of-work • Proof-of-stake • Proof-of-work based derivatives • Federated Byzantines agreement
Centralized validation	<ul style="list-style-type: none"> • Delegated proof-of-stake 	<ul style="list-style-type: none"> • Redundant Byzantine fault Tolerance • Ripple consensus Bilateral node-to-node (N2N) • RAFT and derivatives • Delegated proof-of-stake

- *Block verification*, the recipient verifies the block received in advance. If the block is valid, then the recipient sends the inventory message to the sender. Otherwise, the message is rejected.
- *Inventory message*, it contains information about the block that will be checked by the recipient to ensure the block is never received before.
- *Send getdata message*, this message can be understood as a data request. Afterward the recipient gets the whole information of the block.

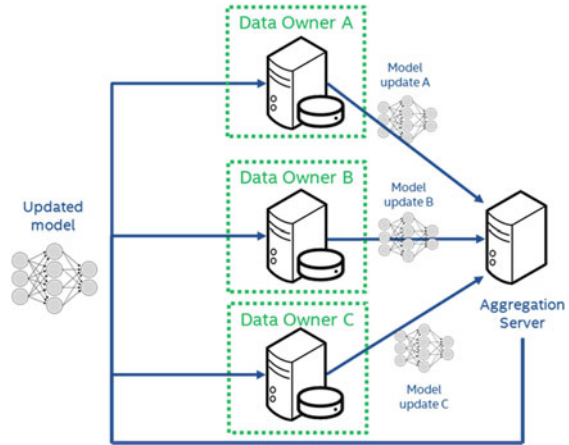
3 Decentralized Machine Learning

One of the most substantial effects of blockchain is not engaging third parties nor intermediaries in managing the transactions. Unlike most centralized systems where every transaction requires intermediaries, the decentralized system is running by the node in the same blockchain network [9]. The principle of a decentralized system can be used in various fields of science, including in the AI area. Large companies with rich datasets control centralized AI. These companies can form a costly data science team to develop models from available datasets. The combination of these technologies has gained much attention due to tremendous advantages over conventional AI method such as efficient in training, and less power consumption.

The merits of decentralized machine learning (on-device machine learning) by leveraging blockchain technology can be summarized as follows:

- *On-device learning (no dataset extraction)*, it protects the privacy of the parties on the network.
- *Multi-blockchain and interoperability*, decentralized machine learning makes it possible to connect with multi-chains to achieve the blockchain-agnostic protocol.
- *Running the algorithm on the owner’s devices*, due to the cryptographic techniques are embedded within the system, the algorithms received can be directly run on the owner’s devices in the same blockchain network.

Fig. 1 Federated learning architecture in general. Each data owner has a local update model that is sent periodically to the aggregation server [10]



- *Federated parties to aggregate the results of AI algorithms.* For instance, by leveraging the federated learning technique [11], as shown in Fig. 1. The encrypted algorithm is broadcasted to the individual institution. Only the updated model is sent to the central model aggregator.

4 The Directions of Decentralized Learning

Research that combines blockchain and decentralized learning has not been much researched because this sort of topic is relatively new. However, since decentralized learning began to be developed by Google recently, research to combine blockchain and decentralized learning began to be widely developed, and startups also emerged.

In this paper, we narrow down decentralized learning to the federated learning model. Decentralized machine learning through a federated model is vital to enable a decentralized group of parties to contribute to the knowledge of an AI model. This protocol is a breakthrough in the AI area that converts centralized raw data into a decentralized form. The conventional machine learning model relies on a centralized form to train the dataset. However, the nature of the centralized model brings up many advantages, such as a single point of failure, and bottleneck issues. Moreover, it has been proven to be challenging in use cases of a large number of endpoints involved in the same model. Therefore, the federated learning technique appears to the public in 2016 as a solution to address the issues.

$$f(w) = \frac{1}{n} \sum_{i=1}^n f_i(w) \tag{1}$$

$$\sum_{n=1}^n \frac{c_n}{c} w_{t+1}^n \tag{2}$$

In the federated learning model, there is a fixed set of clients with a fixed local dataset. For the initial phase, a random fraction of clients is selected by the aggregation server, and afterward, the server sends the updated global model to the clients. Every selected client computes by using their local computation (on-device), and later, the clients send the updated model to the server. The aggregation server then uses these updates as a current global state, and the steps repeat.

Intuitively, $f_i(w) = l(x_i, y_i; w)$ which can be defined as a loss of the prediction on example (x_i, y_i) [12] managed by model parameters w as shown in (1). It also can be interpreted as IoT devices send gradients or parameters to the cloud server $\Delta w^1 + \Delta w^2 + \Delta w^3 + \dots + \Delta w^n$, which is partitioned homogenously. Then, the server aggregates the values and applies it to the new parameters as defined in (2) (Fig. 2).

The decentralization of the federated model is the same as blockchain. Each data owner gets a reward every time they send an updated model to the aggregation server. As in the general, the blockchain has a role in maintaining the consistency of the ledger and in managing revenue generation. We calculate the time needed for a straightforward transaction and the accuracy of the communication. The number of parties involved is five parties, with one aggregation server within ten rounds. Overall, the average time needed for one transaction is 2.82 s, with an accuracy level of 0.89. However, the federated model used affects performance. There are many types of federated models such as vanilla federated learning and its categorization, for instance, horizontal model, vertical, and federated transfer learning.

In order to implement a decentralized AI model in the real world, several challenges must be solved, including:

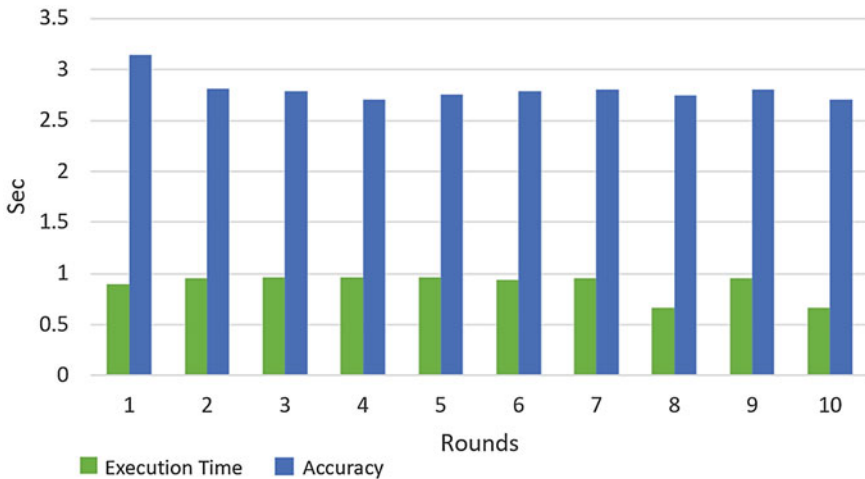


Fig. 2 The results of the federated model over distributed devices in the same network (in terms of execution time and accuracy)

- The privacy issue. The parties can train the AI model without having to reveal their datasets.
- The influence issue. Third-party influences in the knowledge of AI model is also a concern.
- The economic difficulties. Providing incentives to miners and parties involved must be done correctly and fairly to maintain the quality of an AI model.
- The transparency issue. The system must be able to manage every activity in the network that is transparent without the need to trust centralized authority.

5 Conclusion and Prospects

The merits of decentralized learning in the AI area can overcome the issues of the centralized AI system in general. Dependence on third parties can be eliminated, and data owners can contribute directly without having to disseminate their data to the public (privacy concern). Blockchain technology is used to maintain the consistency of the ledger and manage rewards for the parties. The system is tamper-proof, and transaction records are adequately maintained since it is protected by cryptographic protocols. However, some challenges must be addressed before decentralized AI is applied. For an example is the privacy issue of the user. Although the dataset is not available to the public, the addresses used on the blockchain and transaction types can still be seen publicly. Therefore, several cryptographic techniques are suitable to tackle these problems such as ring signature protocol, stealth address, and to name a few.

Acknowledgements This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2018R1D1A1B07048944) and partially was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2015-0-00403) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

References

1. Biais B, Bisière C, Bouvard M, Casamatta C (2019) The blockchain folk theorem. *Rev Financ Stud*
2. Yaga D, Mell P, Roby N, Scarfone K (2018) *Blockchain Technol Overv*
3. Giannakis GB, Ling Q, Mateos G, Schizas ID, Zhu H (2016) Decentralized learning for wireless communications and networking
4. Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning. *ACM Trans Intell Syst Technol*
5. IoT Analytics. <https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/>. Accessed 12 August 2019
6. Rahmadika S, Kweka BJ, Latt CNZ, Rhee KH (2019) A preliminary approach of blockchain technology in supply chain system. In: *IEEE international conference on data mining workshops, ICDMW*

7. Rückeshäuser N (2017) Do we really want blockchain-based accounting? Decentralized consensus as enabler of management override of internal controls. 13. Int Tagung Wirtschaftsinformatik 16–30
8. Rahmadika S, Rhee K.-H. (2019) Toward privacy-preserving shared storage in untrusted blockchain P2P networks. *Wirel Commun Mob Comput*
9. Montes GA, Goertzel B (2019) Distributed, decentralized, and democratized artificial intelligence. *Technol Forecasting Soc Change*
10. Intel AI, Federated Learning Architecture, 19AD. <https://www.intel.ai/federated-learning-for-medical-imaging/#gs.wnvtct>
11. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W (2018) Federated learning of predictive models from federated electronic health records. *Int J Med Inform*
12. Brendan McMahan H, Moore E, Ramage D, Hampson S, Agüera y Arcas B (2017) Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th international conference on artificial intelligence and statistics, AISTATS*

A Study for Accelerating of Convolution Operations Based on Multiple GPUs with MPI



**Boseon Hong, Geunmo Kim, Sungmin Kim, Jeong-Dong Kim,
and Bongjae Kim**

Abstract Advances in hardware and software can handle large-scale data in the field of deep learning. It generally uses a parallel GPU to perform large-scale operations, but there is a limit to the number of GPUs available on a single node. Using multiple nodes can overcome this constraint. In this paper, we designed and evaluated a convolution operation method using the Message Passing Interface (MPI) to accelerate the performance using constant memory which is a kind of the GPU's memory, and multiple nodes with the multiple GPUs environment. According to the performance evaluation results, the use of constant memory of GPU reduces the convolution operation execution time and the convolution operation execution time decreases as the number of GPUs increases in the multiple computing nodes.

Keywords Convolution · Multiple nodes · Multiple GPUs · Message passing interface · CUDA

B. Hong · J.-D. Kim · B. Kim (✉)

Department of Computer and Electronics Convergence Engineering, Sun Moon University,
Asan-si, South Korea
e-mail: bjkim0422@gmail.com

B. Hong

e-mail: goodcools34@gmail.com

J.-D. Kim

e-mail: kjdvhu@gmail.com

G. Kim · S. Kim

Division of Computer Science and Engineering, Sun Moon University, 70, Sunmoon-Ro 221
Beon-Gil, Tangjeong-Myeon, Asan-si, Chungcheongnam-do 31460, South Korea
e-mail: rootmo96@gmail.com

S. Kim

e-mail: kimsungmin.dev@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_19

1 Introduction

As the advancement of the hardware and software, it enables large-scale operation processing. For example, the research area related to deep learning using large-scale data has also advanced. The GPU has made a major contribution to this advance. Generally, a GPU has more computing power than a typical central processing unit (CPU). The GPU has a hardware configuration and architecture different from that of a general CPU. The operating speed of the CPU core is higher than the each GPU core. However, GPUs can do more parallel processing jobs due to a large number of GPU cores. Also, the GPU is composed of onboard memory (global memory) and on-chip memory (shared memory, constant memory, etc.) [1]. On-chip memory, which reads and writes faster than onboard memory, it can be used to customize and optimize computing operations. Therefore, it is more effective to properly use the GPU's on-chip memory than global memory of GPU only to enhance the efficiency of the computing.

Depending on the hardware specification of each computing node, the maximum number of GPUs supported is limited. Using multiple compute nodes allows using more GPUs over the limitation. MPI (Message Passing Interface) is a programming library used to exchange information between compute nodes and is widely used for cluster computing [2–4]. However, the overhead on the message to exchange the results of the operation may increase as the number of computing node increases. Therefore, when designing an algorithm that is operated based on multiple nodes, it is necessary to consider the overhead in terms of data transmission among each computing node. In this paper, we proposed an accelerating scheme of convolution operation which is widely used in the deep learning-based application fields. Our scheme is devised to be effectively operated and applied on multiple nodes with multiple GPUs.

The rest of this paper is organized as follows. In Sect. 2, our scheme for convolution operation is explained. Section 3 describes the experimental environment and the experimental results. In Sect. 4, the experimental results will be explained and discussed. We will conclude this paper with some future works in Sect. 4.

2 Acceleration of Convolution Operation Based on Multiple Nodes and Multiple GPUs with MPI

In order to operate on multiple nodes and multiple GPUs, convolution operations are implemented based on MPI and considering the memory structure of the GPU. There are three existing algorithms for calculating convolution. There is direct convolution, unrolling based convolution, and FFT based convolution [5]. The direct convolutional computation method used in this paper allows the use of constant memory of GPU memory based on multiple nodes and multiple GPUs using MPI.

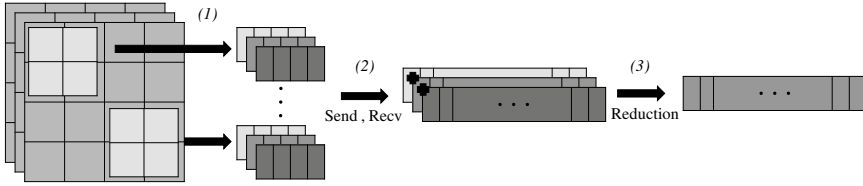


Fig. 1 A concept of convolution computation algorithm based on multiple GPUs and multiple computing nodes

GPU kernel functions use constant memory in on-chip memory to accelerate computational performance. Constant memory is read-only memory. Considering this characteristic and limitation, it was designed and implemented to be used as a filter in convolution. MPI enables data communication with other processes in a multi-node computing environment. The point-to-point communication method is used as the communication method.

Figure 1 shows the procedure of our convolution operation algorithm based on multiple GPUs and multiple nodes. In (1), each thread of the GPU cores performs convolution operation on the size of a filter of one channel. In (2), the result of the convolutional operation is transmitted and received at each rank of MPI environment through point-to-point communication to collect results. Ranks can be thought of as a process, and each rank has a unique value. In (3), the results of the convolution operations collected in the one-dimensional array are reduced according to the number of filters.

3 Performance Evaluation Environment

Table 1 shows the experimental environment to evaluate the proposed accelerating scheme of convolution operation based on multiple GPUs with MPI. As shown in Table 1, each computing node has different CPU specifications. In the case of

Table 1 Cluster computing environment used in the performance evaluation

Features	Computing node 1	Computing node2
CPU	Intel i7-6850 K (6core, 3.6 GHz)	Intel i7-8700 K (6core, 3.7 GHz)
RAM	64 GB (8 × 8 GB)	
GPU	NVIDA GTX 1080Ti × 2	
OS	Ubuntu 16.04	
CUDA versions	CUDA 9.0	
MPI versions	MVAPICH2-2.3	

Table 2 Measurement results of data transfer rates between cluster compute nodes

Features	Description
Network bandwidth	9.43 Mbits/sec

the RAM and GPU, each computing node has the same specifications. The operating system is Ubuntu 16.04 and the CUDA [6] version is 9.0. The MPI version is MVAPICH2-2.3 [7]. Table 2 shows the network bandwidth performance between Computing Node 1 and Computing Node 1. As shown in Table 2, the network bandwidth is 9.43 Mbits/sec.

The experiment was conducted in two parts. First, we experimented about the performance of using main memory which is onboard memory and constant memory which is on-chip memory in a single computing node with a single GPU environment. This is an experiment on memory access efficiency of the memory sharing of each core of GPU by using constant memory. The performance of the convolution operation was measured when the input size was fixed at $1 \times 3 \times 1024 \times 1024$ (meaning: batch size \times channel size \times height \times width). The other experiment is about the performance of convolution operation based on multiple GPUs in multiple computing nodes. In the experiment, we changed the input size and filter size to overload the convolution operation. In all experiments, the stride size was set to 1 and there was no padding value. The number of threads in the GPU was set to 1024. 1024 is the maximum size of the GTX 1080 Ti. The number of blocks of the GPU was set based on Formula 1.

$$\text{Number of Blocks} = (\text{Output Size} \times \text{Input Channel}) / 1024 + 1 \quad (1)$$

4 Evaluation Results

Figure 2 shows the result of the convolution operation. As shown in Fig. 2, the execution time of convolution using constant memory was reduced by up to about 77%, the average reduction was 67% when compared to the global memory only. Therefore, we can confirm that using constant memory is much more efficient than global memory only.

Figure 3 shows the performance results of the convolution operation with multiple GPUs and multiple GPUs. The time spent on convolution operation was reduced by up to 74% when using multiple nodes and multiple GPUs. However, if the filter size and input size are relatively small, using multiple nodes and multiple GPUs is rather inefficient. This is due to that the communication overhead of convolution results sending and receiving between the computing node and multiple GPUs is relatively high. Besides, it can be seen that as the number of GPUs increases linearly, the convolution operation time does not decrease linearly.

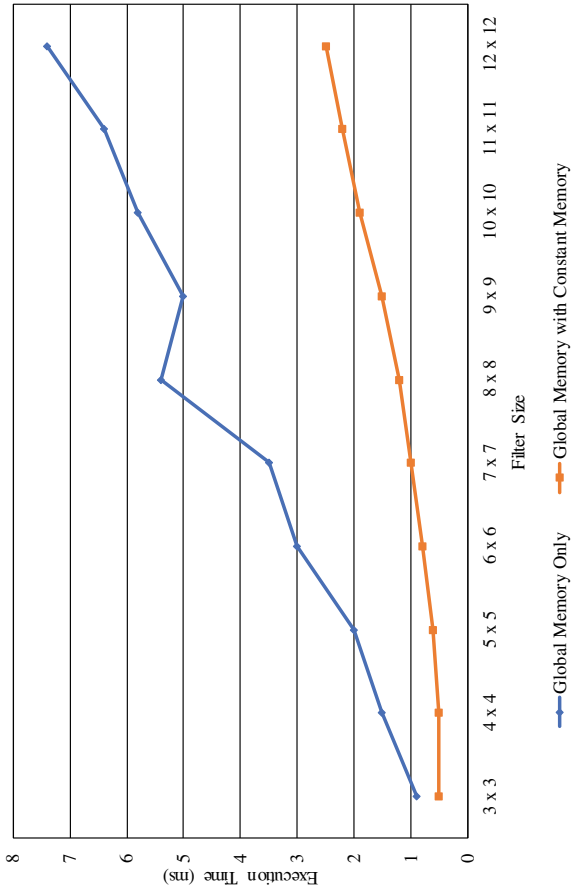


Fig. 2 The result of convolution computation by using constant memory and global memory

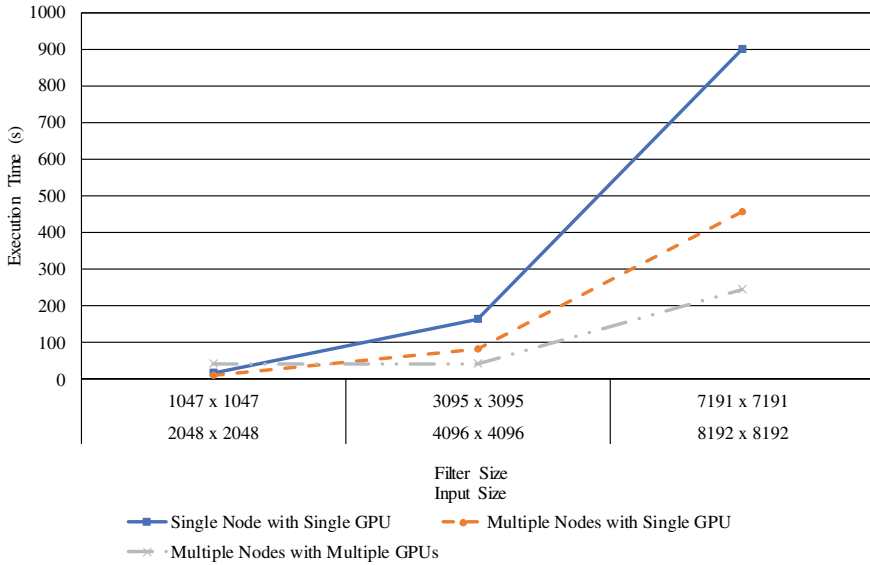


Fig. 3 The performance results of the convolution operation with multiple nodes and multiple GPUs

5 Conclusion and Future Works

In this paper, we proposed an accelerating scheme of convolution operation which can be applied on multiple computing nodes and multiple GPUs by considering the characteristics of GPU architecture. The proposed technique uses not only the global memory of the GPU but also constant memory that operates at high speed. We also performed and analyzed the performance of the proposed scheme. Based on the experimental results, convolution operation using constant memory showed better performance than using the global memory of the GPU only. We also found that if the input data is small, it is inefficient to use multiple nodes and multiple GPUs. However, as the computations covered by the GPU increased, we found that using multiple nodes and multiple GPUs was effective.

GPU's constant memory has the disadvantage that it can only be used to read operation. The shared memory of the GPU is capable of both read and write operations and is faster than global memory. In the future works, we will improve the computational efficiency by additionally applying shared memory of the GPU to overcome this drawback.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2017R1C1B5017476), and supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2018-0-01865).

The corresponding author is Bongjae Kim.

References

1. Nickolls J, Dally WJ (2010) The GPU computing era. *IEEE Micro* 30(2):56–69
2. Kim B, Jung J, Min H, Heo J, Jung H (2017) Performance evaluations of multiple GPUs based on MPI environments. In: *Proceedings of the international conference on research in adaptive and convergent systems*, pp 303–304. ACM
3. Lu J, Zhang K, Chen M, Ma K (2013) Implementation of parallel convolution based on MPI. In: *Proceedings of 2013 3rd international conference on computer science and network technology*, pp 28–31. IEEE
4. Chang LC, El-Araby E, Dang VQ, Dao LH (2014) GPU acceleration of nonlinear diffusion tensor estimation using CUDA and MPI. *Neurocomputing* 135:328–338
5. Li X, Zhang G, Huang HH, Wang Z, Zheng W (2016) Performance analysis of gpu-based convolutional neural networks. In: *2016 45th International conference on parallel processing (ICPP)*, pp 67–76. IEEE
6. Sanders J, Kandrot E (2010) *CUDA by example: an introduction to general-purpose GPU programming*. Addison-wesley professional
7. Panda DK (2013) MVAPICH2: a high performance MPI library for NVIDIA GPU clusters with InfiniBand. In: *GPU technology conference*

Requirements of Future Network for Blockchain Platform Operation



Suyeon Kim

Abstract In this paper, we discuss requirements of Future network to support Blockchain platform. Blockchain is an innovative trading system and which is expected to be used in various fields with stable security function of each node taking part in the system without central server control. However, due to its three functional limitations (excessive traffic generation, lack of distributed control, and unfairness of network resources) presented in this paper, it is hard to provide complete service in the present network. Future network operating Blockchain services should provide an appropriate protocol and network configuration according to Blockchain applications or their service type. In addition, based on the number of the node participated in Blockchain and the amount of data transmitted, it is necessary to provide appropriate QoS and the subnetwork should provide stable protocol function and efficient traffic operation. This Future network structure will be proposed as the future network structure of ISO/IEC JTC1 SC6 being standardized.

Keywords Future network · Blockchain · Blockchain subnetwork · P2P network · ISO/IEC JTC1 SC6

1 Introduction

Blockchain first proposed to implement electronic cryptography called Bitcoin is a system that can be traded without the control of a reliable central server in a P2P based network and is drawing attention as a core technology of the fourth industrial revolution. Recently, with the increase of O2O (Online to Offline) transactions using PC and smart phone, new business models applying Blockchain technology are increasing to overall industry area. Especially Blockchain Technology which cannot be hacked, tampered and counterfeited have been increasing in the Fintech industry.

S. Kim (✉)

Department of Industry Cooperation, Keimyung University, Daegu, South Korea

e-mail: sykim388@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_20

In addition, global financial institutions are pursuing Blockchain system development and standards development through partnership, and Blockchain Platforms are expected to be activated around the world.

However, since the first proposed Blockchain technique was developed as the implementation technique of cryptography, the data structure or protocol method is not universal and is still evolving for deploy. Recently, various kinds of Blockchain platforms have been made to be applicable to other fields. The Blockchain system is expanding into a variety of e-commerce companies including securities companies, banks, trading solution companies, and O2O. It is expected that start-up, virtual money developers, Fintech companies, information protection companies, copyrights, ownership and registration agencies will actively participate in the development and build an ecosystem of Blockchain.

Until now, Blockchain system is operated using an existing network like internet, however it is expected that various Blockchain platforms and application programs different from existing transaction methods will appear in a short period of time. The need for a subnetwork that fits to Blockchain platform and runs it smoothly is expected as shown in Fig. 1.

In this paper, we analysed the problems that can be caused by various Blockchain applications in the current network system. And if we apply these problems to future network requirements that are being standardized in ISO/IEC JTC1 SC6, the optimal future network running Blockchain will be constructed as a new network platform. We think that such a future network will provide users with optimal trading system.

For this purpose, Sect. 2 of this paper briefly introduces the Blockchain system and introduces the problems that can occur in the current network environment. Section 3 presents the requirements of the future network to solve these problems and suggests to apply these requirements to future network standardization in progress. Finally, Sect. 4 presents conclusions on the requirements for future networks and presents future research plans.

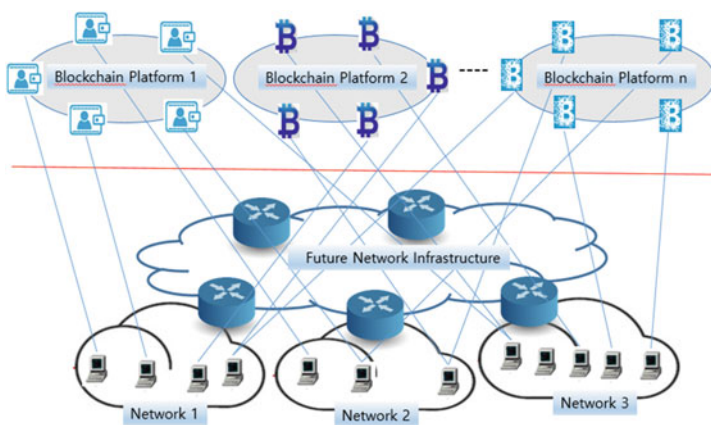


Fig. 1 Blockchain platform over future network

2 Introduction of Blockchain System and Problems at Its Subnetwork

Unlike the conventional method of handling transactions in a centralized server, Blockchain system, core technology of Bitcoin, uses resources (processing power, disk capacity, network bandwidth, etc.) held by participants on a P2P network to which general nodes are connected. This Blockchain technology has emerged as an innovative technology. Bitcoin system based on Blockchain core technology has the same flow chart as Table 1.

In Table 1, the creation of an account in step A is action for creating an individual account in the same way as a bank account. At this time, a private key and a public key are generated. Transactions in step B can be transmitted or received in Bitcoin data using the public key as the receiving address (or account number). The data could represent money, contract, deeds, medical record, customer details, or any other asset that can be described in digital form. The transaction verification of step C is performed by transmitting the transaction generated in step B to the neighbouring nodes. If the transaction is judged to be appropriate through this verification process, it is propagated again, and if it is judged that the transaction is not appropriate, the transaction is deleted. The transaction is either verified instantly or transcribed into a secured record and placed in a queue of pending transactions.

The block configuration and creation of step D stores the transaction judged appropriate by each node in the memory pool for block production. Each block is identified by a hash, a 256-bit number, created using an algorithm agreed upon by the network. The mining and compensation of step E is the process of solving the block by making incremental changes to one variable until the solution satisfies network-wide target. Block validation of step F confirms the appropriateness of the block from the neighbour node by transferring the generated block to the neighbour node.

Table 1 Overall flowchart of bitcoin transaction

Life cycle	Action
A. Account creation	Electronic wallet creation
B. Transaction creation	Bitcoin exchange
C. Transaction verification	Transaction data broadcasted to P2P network
D. Block configuration and creation	Transaction combined to form a data block
E. Mining and compensation	A block contains a header, a hash and a group of transaction
F. Block validation	Check hash function and Select majority blockchain
G. Blockchain creation	Adds the block to the majority blockchain
H. Estimated difficulty retarget	Bitcoin hash rate

If more than half of all nodes agree with the appropriateness of the generated block, it is approved as a new block.

In the step G, the node that receives the block generated in step F completes the verification of the convex and connects to the existing block chain. Because of the distributed data structure, the mined blocks can be created at the same time from other nodes, which can result in Blockchain branching (Forking) due to the propagation delay of the network. In fact, in the Bitcoin, only the longest block chain survives and the remaining chains disappear into short orphan blocks. The degree of difficulty of step H adjusts the probability of self-occurrence that difficulty in mining so that the block generation cycle can be adjusted every 10 min.

Since the Blockchain is a distributed data structure, the mined blocks can be simultaneously generated at different nodes at the same time. Such blocks may be transmitted to the nodes participating in the platform at the same time, resulting in a large number of branching phenomena. An example of an agreement algorithm for solving the branch phenomenon is 'The Byzantine Generals Problem' (BGP). BGP solution is to remove these orphaned blocks if less than one-third of the nodes present an orphan block [1].

As the branching phenomenon occurs more frequently, the number of orphan blocks not connected to the chain is increased, resulting in network congestion and a waste of bandwidth. In the Blockchain system, since only the longest block chain survives, the slower the network processing speed, the more often the orphan block is generated.

How many blocks to process per second is important for Blockchain performance. However, the Blockchain performance such as the finality of each transaction, the cycle in which blocks are generated, the amount of processing variation depending on the block size, the number of nodes showing decentralization, the efficiency of the algorithm, and parallel scalability, should be considered together [2].

In this way, I am studying on Blockchain from a network perspective, i.e., how information (transaction and blocks) is disseminated or propagated in the network. However, the problems that may occur in the subnetwork due to the Blockchain system cannot be overlooked. In this paper, we are going to bring up the following three problems.

- (1) Possibility of excessive traffic on the subnetwork side: Mass traffic generated in the process of blocks, transactions and validations caused by various Blockchain applications, and large amounts of traffic due to frequent data exchange for block validation process through block transmission and consensus process to all participants in the network.
- (2) Absence of decentralized control: Subscription and transaction of Blockchain can be done by anyone, and some nodes can generate large amount of data locally. Because of this problem, it is impossible to control the traffic load in the Blockchain system overwhelming network bandwidth.
- (3) Fairness of network resources: In Blockchain system, a node with a lot of computing resources monopolizes block generation, resulting in unfairness in network bandwidth usage due to unbalanced computing resources.

3 Requirements of Future Network to Support Blockchain System

In addition to the problems that the Blockchain system mentioned in Chap. 2 may cause in the subnetwork, there may be more problems. However, in this paper, we consider the requirements of the future network focusing on the above three problems.

In ISO/IEC JTC1 SC6, we are already studying the subnetwork to support various application programs including Blockchain, and its standardization is in progress taking into account the expected requirements. The details of the future network requirement are defined in the document TR 29181. Based on this document, standardization of ISO/IEC 21558: future network structure and ISO/IEC 21559: future network protocol is currently underway [3, 4]. ISO/IEC 21558 and ISO/IEC 21559 Future network documents are subdivided into areas as shown in Table 2.

For smooth operation of the Blockchain system, the requirements of the future network belonging to the lower layer are analysed as follows.

- (1) **Necessity of Traffic Control:** It is necessary to control traffic by measuring the number of the blocks occurring in each node and to adjust it so that excessive traffic does not occur in the network. In order to adjust the traffic, we need to study and enhance the routing protocol of the future network considering the tree configuration for verifying the chain validity of the Blockchain. This work is related to the routing configuration of ISO/IEC 21558-2 and ISO/IEC 21559-5.
- (2) To construct an effective Blockchain system, various subnetwork configurations need to be constructed for each Blockchain platform, and subnetwork should support these configurations. For this purpose, the network is configured to minimize the generation of orphaned blocks and the network can be reconfigured if a large number of orphaned blocks are created.
- (3) The subnetwork should grasp the usage of computing power and network resources in a legitimate way, controls nodes that use many resources, and equally distribute resources to maintain fairness of network traffic. If the particular node uses excessive traffic on the specific platform, the subnetwork has to adjust the value of QoS in ISO/IEC 21558-3 and ISO/IEC 21559-6 to control the traffic usage in proportion to the cost.

Table 2 Future network standardization area

Document number	Topic
ISO/IEC 21558-2	Architecture part 2: switching and routing
ISO/IEC 21558-3	Architecture part 3: quality of service
ISO/IEC 21558-4	Architecture part 4: network of everything
ISO/IEC 21559-5	Protocol part 5: switching and routing
ISO/IEC 21559-6	Protocol part 6: quality of service
ISO/IEC 21559-7	Protocol part 7: network of everything

- (4) For the smooth operation of the Blockchain system, the subnetwork needs to provide stability and robustness. As Blockchain platforms become more diverse and the number of transactions is expected to increase exponentially, the subnetwork must be stable independent to the amount of traffic they generate. A number of nodes participate in the Naming scheme of NoE (Network of Everything) and a lower network layer should be configured to facilitate traffic management by the switching scheme of NoE.

4 Conclusion

Blockchain platforms are expanding into securities companies, banks, trading solution companies, and e-commerce companies including O2O. It is expected that various development agencies will participate in building a new ecosystem for Blockchain. In this paper, we study on the necessary requirements of the future network as a subnetwork structure to provide stable protocol function and efficient traffic operation of several Blockchain Platforms.

After analysing the problems that may arise from Blockchain applications in the current network system such as how information in the Blockchain is disseminated in order to synchronize the information of each node and how Blockchain fork occurs, we introduced the requirements to solve these problems in the Future Network standardized by ISO/IEC JTC1 SC6.

Based on the requirements presented in this paper, we will define the services of the Blockchain platform over the Future Network [5]. I believe that standardizing the protocol mechanisms to provide such services is now a matter of preparation. If this standard work is completed, the Blockchain and the optimal Future subnetwork will be able to provide the optimal trading system to the users as a new Blockchain infra-structure.

As considering the fast processing speed that will be the basis of the 5G network or the future network, the generation of orphan blocks may be minimized according to the network configuration method in the long term. And if you have to charge a fee based on network usage, you can simplify the algorithm of the Blockchain or you can save the network usage fee and computing power by managing the network bandwidth. This changes may mitigate the problem of Blockchain fork in the long term.

Acknowledgements This paper was prepared by the support of the ‘National Standard Technology Improvement Project’ of the Ministry of Commerce, Industry and Energy (Project Number: 20002532).

References

1. Nagato K, Wataru S (2018) Blockchain application kaihatsu no kyokasho. Mynavi Publishing Corporation, Japan
2. Bitcoin developer reference. <https://bitcoin.org/en/developer-reference>
3. ISO/IEC 21558-x—information technology—Telecommunications and information exchange between systems—Future network architecture
4. ISO/IEC 21559-x—Information technology—Telecommunications and information exchange between systems—Future network protocols and mechanisms
5. ISO/IEC JTC1/SC6 N16860: summary of voting on the establishment of an ad-hoc group on networking for blockchain and its ToR on 6N16841
6. Kim W (2018) Bitcoin blockchain operation principle and evolution (in Korean). Weekly ICT Trends 1851:2–15
7. Lim M (2016) Analysis of utilization trend of blockchain technology (in Korean). Weekly ICT Trends 1772:2–14
8. Lim M (2006) ZIB structure prediction pipeline: composing a complex biological workflow through web services. In Nagel WE, Walter WV, Lehner W (eds) Euro-Par 2006. LNCS, vol. 4128. Springer, Heidelberg, pp 1148--1158
9. Satoshi N, Bitcoin: a peer-to-peer electric cash system. <https://bitcoin.org/bitcoin.pdf>
10. Kang SG, Hyung W, Lee CK, Proposed updates to study report of PWI-P2P, Functional architecture and protocols for managed P2P communications. ISO/IEC JTC1/SC6/WG7 N154.

TELL ME: Design of an Intelligence-Empowered Recommendation System



Kuan-Hua Lai, Neil Y. Yen, and Jason C. Hung

Abstract With the advances in science and technology, the life becomes smarter and more convenient. In recent years, the Artificial intelligence (AI) is a popular topic. AI can be used in many areas such as healthcare, Automotive, Finance, etc. For instance, the smart phone could be a multi-function entity such as voice connecting, data storing, information displaying, etc. These are expected to provide more significant and precise function for each user based on personal requirement. This study, with the concern on high-quality service provision, seeks for the possibility in developing location-based service(s) based on the concept of anticipatory computing. We attempt to achieve the concept of request-free and realize the scenario of ‘act before you act.’ Hence, three main portions will be discussed.

Keywords Anticipatory computing · Location-based service · User understanding · Personality model · Cognitive-Behavior model

1 Introduction

Nowadays, use of virtual assistant, that serves individual’s needs, becomes more and more popular. The smart phone become more convenience than before that can provide lots of services, for example user can use virtual assistant to set alarm, search weather, search information, etc. Nowadays, there are lots of virtual assistant such as, Siri, Alisa, Amazon Alexa, Google Assistant, etc. TELL ME is a recommendation system based on virtual assistant. The purpose of this research TELL ME is

K.-H. Lai · N. Y. Yen

School of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu, Japan
e-mail: frank12690@gmail.com

N. Y. Yen

e-mail: neilyyen@u-aizu.ac.jp

J. C. Hung (✉)

Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taichung, Taiwan
e-mail: jhungc.hung@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_21

recommended something to the user at the specific time and location. This is need to analyzed users' past behavior. Based on users' past behavior to get the user personality traits data that based on Big-Five personality traits model. The cognitive model combined Big-Five personality traits to build a user model. The user to do some things that will be get feedback from the user that is state transformation. The state transformation with the decision processes have two type. The first type is complete situation to decision that is user follow the TELL ME recommendation. The second type is incomplete situation to decision that is user no follow TELL ME recommendation to decision. As mentioned above, in the traveling area, TELL ME can base on user location, time to recommend what can user do. TELL ME is an integration of three aspects, the temporal, the spatial, and personality traits.

2 Related Work

2.1 Mobile Behavior Change User Life

Nowadays, a lot of people use smartphone. The smartphone changes people's life. The smartphone is powerful, moreover includes powerful function of handling information, a good of interaction interface and an easy operating of sensors such as light, location, proximity, and acceleration sensors. The smartphone become more and more convenient and smart. It is completely change users' life.

The smartphone can automatically monitor users through sensors. Furthermore, smartphones can use sensors to guess the user's physical activity, movements, stress, and emotional states. This is can predict users' behavior.

For example, the Bewell use smartphone monitoring to record user activities such as sleeping, exercising and socialization based on sensors [1]. Bewell using wellbeing model user behavior patterns and estimate wellbeing score. Moreover, it can design users' health situation. Hypothetically, the user changes the behavior, the sensors will tell the user how to improve.

2.2 Application Predictive User Behavior

The user can be recommended, and reminder based on location-predicting. For example, the application reminds the user need to pack up the book when user on a day past the library. The Magitti recommendation system have user and other's past activities record, as well as can recommend other user in the correct location to do same things [2]. The predictive ability, such as Magitti recommendation system can recommended some activities to user a specific point in time and remind user need to prepare something [5].

There are some predictive applications based on location, such as Global Positioning System (GPS) navigation application can guide user when user miss a turn or road. The “opportunity knocks” systems can remind user when user take the bus to his/her destination [3]. In other words, the user takes the wrong bus or misses the bus stop, in addition the application can guide the user to correct route. The application has a feature is reminded the user need to get off the bus.

2.3 Spatial-Temporal Prediction

The spatial-temporal prediction (STP) can apply in our life such as the user will appear in a specific time or location. The prediction result can show to user what the user did before. Meanwhile, the system can recommend what can user do at that time. The social media base on spatial-temporal to recommend the similar user personal traits people to. In addition, the user can easily connect to other users. In our research, based on spatial-temporal to predict where the user shows up and what can user do [4].

3 Proposed Approach

The TELL ME is an intelligence-empowered recommendation system. The scenario structure as show in Fig. 1. The system will crawl the data from cyber world and

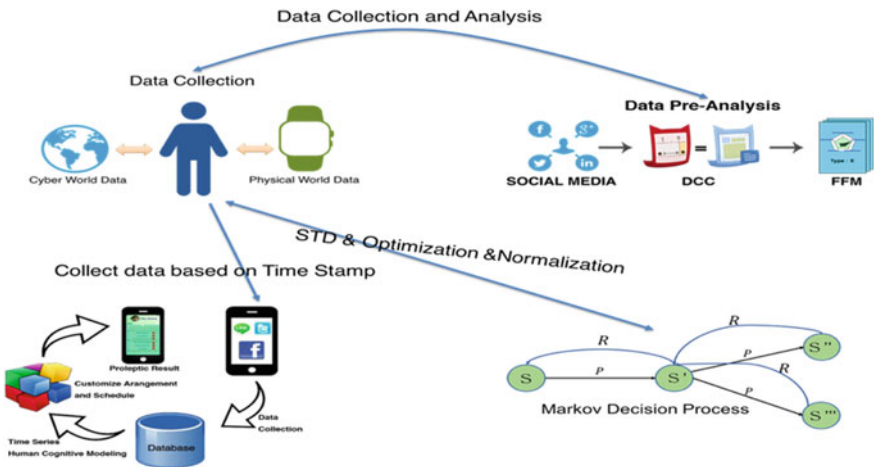


Fig. 1 TELL-ME structure

physical world. In addition, the users' wearable devices will provide the environmental context to system. The system will analysis users' data, moreover, the system will suggest the prediction result to user. There are four steps in TELL ME.

I. Data collection

In this research, the data sources are from Cyber data and Physical data. The Cyber data is user in virtual world behavior such as surf the internet, use the social media, and read E-book etc., on the online internet. The Physical data is user do some behavior in real world such as heartbeat/blood-pressure, moving distance, and location etc.

II. Data Pre-Analysis

The data pre-analysis after data collection. In this step, we will create user profiling model. The user may have different background, different habit, and different ability etc. In addition, we can have based on user profiling model to do personalized analysis. For example, the user personality traits are happy and humanize, furthermore, the user can accept the positive news higher than negative news.

III. Post-Data Processing

There are three calculation process in post-data processing. The post-data processing calls Anticipatory Model. First, this research has modified and optimize the Markov Decision Process (MDP). In addition, using MDP to get predictive result, user selection, and the feedback of state transfer process. Second, according to Theory of Planned Behavior (TPB) lead to a serial feedback result about user input the different information. Based on TPB to modify the MDP. Third, this research has designed a new method about user expected feedback of tolerance interval. This method is description the user's good or bad selection in specific situation. Moreover, the relevance of possible feedback.

IV. Result display

Finally, this research designs an application about user oriented. This application focus on expected results based on past models. This application design style is simply and directly. This design style makes the user obtain the information easily and clearly. At the same time, get the feedback immediately. Additionally, putting the social media concept in an application. The users calculate by himself/herself, he/she also can switch to another users. To browse another user's expected result in the same situation. This is can improve users' interactive.

In this research, modified some techniques. Moreover, developed some algorithms to achieve the expected outcome.

4 Method

In this research, we aim to predict user behavior that we design two models. So that we through cognitive action model to design user personality traits.

1. User Cognition Action Model

We will through Theory of Reasoned Action (TRA) to discuss and analysis. This theoretical model is mainly from the perspective of psychology. Through four main variables: belief, attitude, intention, and behavior, explain why the user takes action in a particular situation. At the same time, through this process to further speculate on the user’s action pattern as shown in Fig. 2.

Moreover, we design a function for user to do pre-calculate. The first is to explain the relationship between the user’s intention (I) and the actual action (B) of the specific behavior. The user’s behavioral intention is affected by two factors, one is the behavioral attitude (A_B). The other is social pressure (SN) outside the user who takes the action.

$$B \sim I = (A_B)w_x + (SN)w_y \tag{1}$$

B is the user’s code of conduct, I is the user’s intention for behavior B , AB is the user’s attitude toward behavior B , and SN is the user’s social subjective norm for behavior B . Among them, we can deeply explore the user’s attitude towards a certain behavior, and this attitude will be revised due to the change of the user’s current thoughts (or beliefs).

$$A_B = \sum b_i e_i \tag{2}$$

B_i is the i result produced after taking action B , and e_i is the evaluation of the feedback for the selected i . Finally, it is the subjective normative factor of the external environment for the user. We take into account the expectations of the factors such

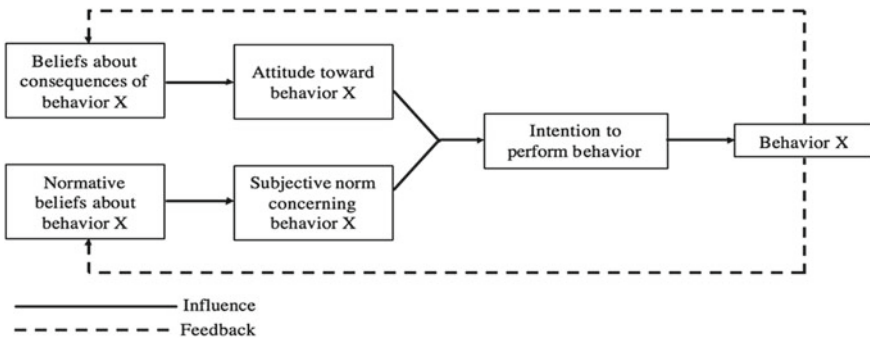


Fig. 2 Cognitive-action model diagram

as, specific people or groups that the user wants to be specific behaviors. And the user tries to satisfy this expectation.

$$SN = \sum b_j m_j \quad (3)$$

B_j is the number of expectations that the user believes that a particular external factor j is taking action B , m_j is an attempt to obey j , and m is expected.

5 Result

In the result, we aim to find the suitable house to guest. So, we need to know the user personality traits and the user thinking. The user personality traits method is cloning the IBM Personality Insights. This method is personality traits based on text analysis. The personality traits can largely apply in several areas. The data source is based on Airbnb. We have written a crawler to get data.

The analyze result as shown in Fig. 3. In addition, there are three primary models in visualization result. The blue area is Big Five personality traits that includes Agreeableness, Conscientiousness, Extraversion, Emotional range, and Openness. The comments show the guest personality trait is Openness. Moreover, the guest is energetic: he/she enjoys a fast-paced, busy schedule with many activities. He/she is self-assured: he/she feels he/she have the ability to succeed in the tasks he/she set out to do. The green area is Needs which meaning is aspects of a product are likely to resonate with a person. The guest's Needs result is Closeness which meaning is being connected to family and setting up home. The red area is Values which meaning is motivation factors that influence a person's decision making. The guest's Value result is Conservation which meaning is emphasizing self-restriction, order, and resistance to change.

6 Conclusion

TELL ME is a recommendation system based on virtual assistant. Moreover, this is based on personality traits recommendation system. The recommendation system can provide users with suitable and expected needs in various fields. This study is looking forward to the development of a set of advanced technologies that can be fully utilized. Modules and methods that the next generation of intelligent computing can more effectively achieve the purpose and use in more fields.

The reference of the user behavior and the design of the recommendation system have been researched and developed in recent years. Whether it is in a search engine such as Google or a social network such as Facebook's push system, it is often found that the user's suggestions are often accurately marketed. This will further increase

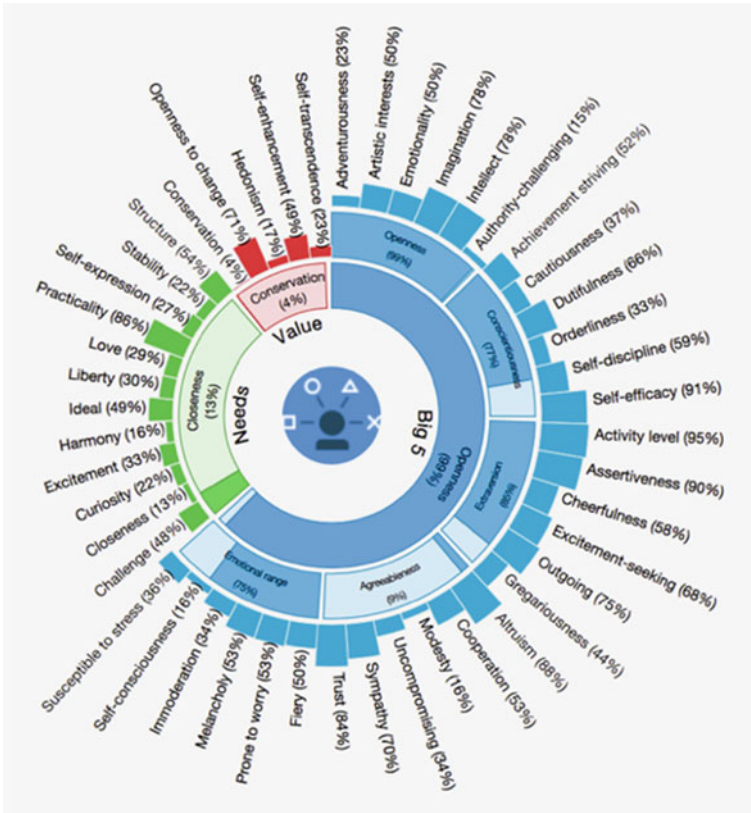


Fig. 3 The visualization results of the user personality traits

user loyalty. Nowadays, in the mobile system, the combination of the Internet of Things and wearable devices is booming. These related applications must be able to meet the needs of users.

References

1. Lane ND, Mohammad M, Lin M, Yang X, Lu H, Ali S, ... Campbell A (2011) Bewell: a smartphone application to monitor, model and promote wellbeing. In: 5th International ICST conference on pervasive computing technologies for healthcare, pp 23–26
2. Burbey I, Martin TL (2012) A survey on predicting personal mobility. *Int J Pervasive Comput Commun* 8(1):5–22
3. Patterson DJ, Liao L, Gajos K, Collier M, Livic N, Olson K, ... Kautz H (2004) Opportunity knocks: a system to provide cognitive assistance with transportation services. In: International conference on ubiquitous computing, pp 433–450. Springer, Berlin, Heidelberg
4. Yang G, Züfle A (2017) Spatio-temporal prediction of social connections. In: Proceedings of the 4th International ACM workshop on managing and mining enriched geo-spatial data, p 6.

ACM

5. Lai KH, Yen NY, Hung, JC (2018) User modeling based on sentiment analysis. In: 2018 9th International conference on information technology in medicine and education (ITME), pp 1072–1076. IEEE

Estimation of Weights in Growth Stages of Onions Using Statistical Regression Models and Deep Learning Algorithm



Wanhyun Cho, Junki Kim, Myung-Hwan Na, Sangkyoon Kim,
and Hyejin Lee

Abstract Generally, the problem of predicting yields of onions grown in a year is of utmost concern to both the farmers who grow vegetables and the government departments that manage them. In this study, we first considered five environmental variables, Mean Wind Speed (MWS), Mean Temperature (MT), Mean Ground Temperature (MGT), Mean Humidity (MH), Daily Sunshine (DS) and Daily Rainfall (DR), which have high influence on onion weight at different stages of growth. Second, we use the partial least square (PLS) regression, support vector machine (SVM) regression, multilayer perceptron (MLP) network as statistical prediction model and LSTM network as deep learning algorithm in order to predict the weight of onion using the collected data. Third, we conducted an experiment to investigate the performance of four prediction models for its weight and the influence of six environmental variables on onion growth. Finally, from the experimental results, we first note that the optimal cultivation strategy to increase onion growth is to lower the MWS, MGT and DR below a certain level and at the same time increase the MT, MS and DS values above a certain level. Secondly, we note that for raw data, the weight of onions is not well predicted at the stages of growth by four prediction methods, but for log transform data, it is well predicted during the growth stages. Thirdly, we can also see that the

W. Cho · J. Kim · M.-H. Na (✉)

Department of Statistics, Chonnam National University, Gwangju 61186, South Korea
e-mail: nmh@chonnam.ac.kr

W. Cho

e-mail: whcho@chonnam.ac.kr

J. Kim

e-mail: kjkwnsrl@gmail.com

S. Kim

Department of Electronic Engineering, Mokpo National University, Jeonnam 58554, South Korea
e-mail: narciss76@mokpo.ac.kr

H. Lee

Rural Development Administration, 300, Nongsaengmyeong-ro, Jeonju-si, South Korea
e-mail: lhj5157@korea.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_22

SVMR method is slightly more predictive than the other three methods, PLS, MLP, and LSTM for both raw data and transformation data.

Keywords Estimation of onion weight · Raw and log transform data · Environmental variables · Statistical regression models · Deep learning algorithm · PLSR · SVMR · MLP network · LSTM network · Accuracy measure

1 Introduction

Generally, governments and agencies want to know how much harvest of Chinese cabbage or onions this year is harvested. Because of the principle of supply and demand of agricultural products, if too much is produced, the price of agricultural products is lowered. On the contrary, if too little is produced, the price of agricultural product will increase sharply. Therefore, if we can know about the yield of agricultural products this year, the government and related organizations can prepare appropriate policies for price control.

On the other hand, how much of the year's yield depends on how farmers cultivated vegetables during the year. Farmers are therefore interested in tillage techniques that can improve yields. Recently, with the development of smart farm technology, various fields such as supply of water and fertilizer and control of pests and diseases are being implemented automatically in farm cultivation [1, 2]. In addition, the related organizations collect data on the environmental factors and growth factors of vegetables daily through various farms in order to check the growth status of the vegetables [3, 4]. However, suitable tillage technique that can perfectly improve the yield of vegetables in Korea has not yet been fully realized.

In this paper, we consider a statistical prediction problem that can automatically predict the development of onions grown in open fields. In general, the yield of onions depends on the state of development in the cultivation process, and this development is also determined by environmental factors or soil levels. Therefore, we are going to develop a prediction system to understand how various environmental factors such as temperature, sunlight or moisture, affect onion weight.

2 Dataset and Methods

2.1 Dataset

In this study, we used datasets of the various environmental factors and the weight of growth stages of onion corrected from farmers in several regions of Korea during from March 2019 to June 2019. We used six factors such as mean wind speed, mean temperature, mean ground temperature, mean humidity, daily sunshine, and daily rainfall as the explanatory variables of the model, and the weight of onion growth

Table 1 The environmental variables used in data set

Role	Variable name	Time lag	Unit of measure
Response variable	Weight of onion	7 days	g/1 unit
Environmental variables	Mean wind speed	7 days	mps
	Mean temperature	7 days	°C
	Mean-ground temperature	7 days	°C
	Mean humidity	7 days	%
	Daily sunshine	7 days	h
	Daily rainfall	7 days	mm

stage was used as a response variable. A description of each variable is given in Table 1. We collected a total of 178 data from 28 farms and used them in the experiment.

2.2 Method

Here we briefly review their algorithms for PLSR, SVMR, MLP and LSTM, which are the methods used to analyse the collected data. First, the Partial Least Square Regression (PLSR), like principal component regression analysis, is designed to solve the problem of going through a large number of highly correlated independent variables. Second, the objective of the least-squares SVM regression method is to map this into a high-dimensional nonlinear feature space using the appropriate transform function $\phi(\mathbf{x}) : \mathbb{R}^k \rightarrow \mathbb{R}^l$ given the observation \mathbf{x} . And in this feature space, we try to predict the estimated value $\hat{y}(\mathbf{x})$ of the response variable using the following linear function.

$$\hat{y}(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$$

where b represents the y-intercept and \mathbf{w} represents the vector of regression coefficients in the nonlinear feature space. Third, the Multi-layer perceptron (MLP) networks are robust in dealing with the ambiguous data and the kind of problems that require the interpolation of large amounts of data. The MLP network consists of three layers, which are input, hidden and output. Nodes also known as neurons are present in each layer. Activation function is in hidden layer and output layer. Fourth, Long-Short Term Memory (LSTM) is an extension of Recurrent Neural Network (RNN) which has not only the recurrent learning unit inside the network but also several gates to capture the longer states from the beginning unit and the shorter

states from the last unit. By having this feature, LSTM has been broadly used to solve time series forecasting problems.

3 Experimental Results

3.1 Accuracy Measures for Model Performance

We adopted the following three measures to evaluate the prediction accuracy of onion weights for four methods. They are the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE), which are defined as follows:

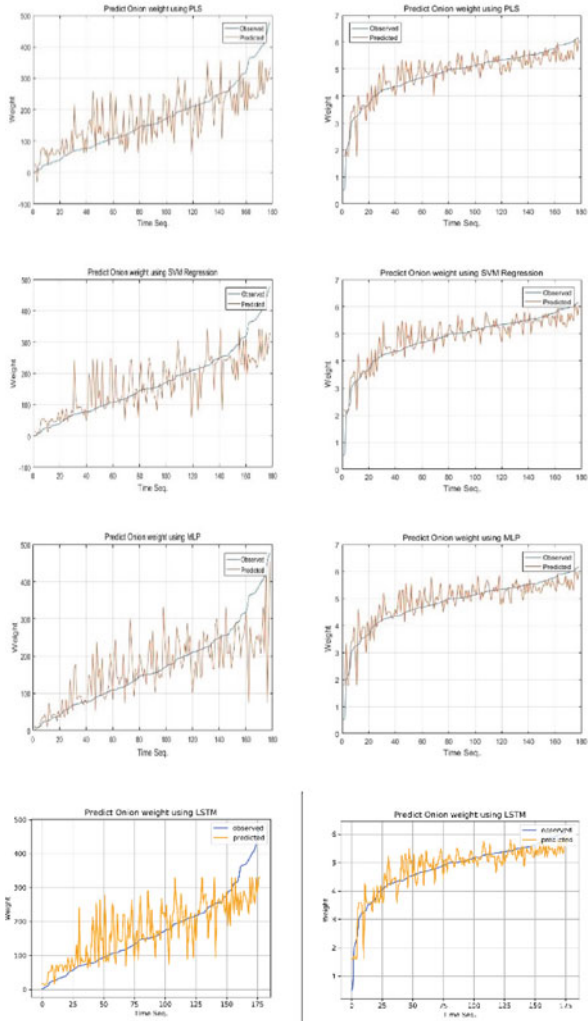
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}.$$

3.2 Performance Evaluations of Four Methods

In order to examine the performance of the four methods considered, we experimented with two cases of raw data of onion weight and its log transformation data. Figure 1 shows a two-dimensional plot of actual observations and predicted values of onion weights at the growth stages by using the four methods such as PLSR, SVMR, MLP and LSTM. The left side shows a two-dimensional plot of raw data and the right side shows a two-dimensional plot of log transformation data. From these graphs, we note that for raw data, the weight of onions is not well predicted at the stages of growth by four prediction methods, but for log transform data, it is well predicted during the growth stages.

We have calculated the measure of the RMSE, MAE and MAPE defined above to compare the performance of the four methods of PLSR, SVMR, MLP and LSTM that predict the weight of onion considered so far. Table 2 shows the values of the error rates by the four prediction methods using raw data and log transformation data. From the results of this table, we can see that the result of applying the log transformation data rather than the raw data reduces the error rate significantly. In addition, it can be seen that the SVMR method lowers the error rate in comparison with other three methods, PLS, MLP, and LSTM.

Next, we calculated the beta coefficients through PLS regression to see how each environmental variable affects the weight of onion growth stage. Table 3 shows the beta coefficient values of each environmental variable over time. From these results, we can see how each environmental variable affects the weight of onions. First,



(a) Raw data for weight (b) Log transformation data for weight

Fig. 1 Prediction of onion weight for PLSR, SVMR, MLP, LSTM

Table 2 Error rates for PLSR, SVMR, MLP, LSTM using raw data and log transformation data

	RMSE		MAE		MAPE	
	RD	LTD	RD	LTD	RD	LTD
PLS regression	70.24	0.386	54.04	0.297	0.647	0.087
SVM regression	69.19	0.406	47.94	0.283	0.436	0.092
MLP network	74.10	0.441	49.37	0.329	0.375	0.098
LSTM	70.08	0.409	51.92	0.318	0.530	0.090

RD Raw Data; *LTD* Log Transformation Data

Table 3 Beta coefficients of environment variables

Period	MWS	MT	MGT	MH	DS	DR
t-7	-7.69445	15.76528	-8.23461	5.838677	2.063574	-0.64000
t-6	-0.60166	-10.1232	31.97899	2.617675	12.43185	-1.65507
t-5	-20.6856	-16.4733	-20.9158	-8.64566	-16.8606	-0.05359
t-4	-41.7138	10.28594	1.187904	-1.00089	-8.01451	-19.7768
t-3	96.57724	-19.5029	3.649576	0.360455	-31.2989	-26.3721
t-2	-54.5091	-9.2566	32.37667	3.591376	1.716278	-0.62559
t-1	-5.27819	12.84363	-7.21797	3.325923	9.096893	-1.62334

MWS Mean Wind Speed; *MT* Mean Temperature; *MGT* Mean Ground Temperature; *MH* Mean Humidity; *DS* Daily Sunshine; *DR* Daily Rainfall

there is a negative correlation between the values of MWS (Mean Wind Speed), MGT (Mean Ground Temperature) and DR (Daily Rainfall) and the weight of onion. As these values increase, the weight of onion tends to decrease. In contrast, MT (Mean Temperature), MH (Mean Humidity), and DS (Daily Sunshine) have positive correlations with onion weight, and as these values increase, onion weight tends to increase. Therefore, the optimal cultivation strategy to increase onion growth is to lower the MWS, MGT and DR below a certain level and at the same time increase the MT, MS and DS values above a certain level. It can be seen that this is given as a cultivation strategy to maximize the weight of the onion.

4 Conclusions

Here, we considered the problem of predicting the weight of onion at each stage of growth that finally affects the yield of vegetables using traditional prediction methods and deep learning algorithm. Experiments were carried out to evaluate a performance of three prediction methods such as PLS regression, SVM regression, MLP network and a deep learning algorithm-LSTM.

From the experimental results, we first note that the optimal cultivation strategy to increase onion growth is to lower the MWS, MGT and DR below a certain level and at the same time increase the MT, MS and DS values above a certain level. Secondly, we note that for raw data, the weight of onions is not well predicted at the stages of growth by four prediction methods, but for log transform data, it is well predicted during the growth stages. Thirdly, we can also see that the SVMR method is slightly more predictive than the other three methods, PLS, MLP, and LSTM for both raw data and transformation data.

Acknowledgements This work was partially supported by the Research Program of Rural Development Administration (Project No. PJ0138672019) and the Korea National Research Foundation (Project No. 2017R1D1A1B03028808) of Korea Grant funded by the Korean Government.

References

1. Zeng W, Xu C, Gang Z, Wu J, Huang J (2018) Estimation of sunflower seed yield using partial least squares regression and artificial neural network models. *Pedosphere* 28(5):764–774
2. Gandhi N, Petkar O, Armstrong LJ, Tripathy AK (2016) Rice crop yield prediction in India using support vector machines. In: Proceedings of 13th international joint conference on computer science and software engineering (JCSSE), July 2016
3. Niedbała G (2019) Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield. *J Integr Agric* 18(1):54–61
4. Jiang Z, Liu C, Hendricks NP, Ganapathysubramanian B, Hayes DJ, Sarkar S (2018) Predicting county level corn yields using deep long short term memory models. [arXiv:1805.12044](https://arxiv.org/abs/1805.12044)

Dynamic Projection Mapping Based on the Performer's Silhouette



Injae Jo, Youjin Koh, Taewon Kim, Sang-Joon Kim, Gooman Park,
and Yoo-Joo Choi

Abstract This paper presents a novel dynamic projection mapping which tracks a silhouette of a stage performer and projects a video image on the tracked silhouette region. In the proposed method, the printed calibration board is not required for the camera-projector calibration. It allows the simple camera-projector calibration for the wide space is possible. The image can be accurately projected on the body of the performer moving freely on the stage due to the calibration considering the different distance from the camera. In the experiments, the video images were projected onto the body of the performer moving freely, not only from left/right to right/left but also from front/back to the back/front of the stage.

Keywords Dynamic projection mapping · Camera-projector calibration · Body projection mapping

I. Jo · Y. Koh · T. Kim · Y.-J. Choi (✉)

Department of Newmedia, Seoul Media Institute of Technology, 661 Deunchon-dong,
Gangseo-gu, Seoul, Korea
e-mail: yjchoi@smit.ac.kr

I. Jo

e-mail: injae1028@gmail.com

Y. Koh

e-mail: ssummerr8.8@gmail.com

T. Kim

e-mail: wingtgniw@naver.com

S.-J. Kim

Department of Information Technology and Media Engineering, Seoul National University of
Science and Technology, 232, Gongneung-ro, Nowon-gu, Seoul, Korea
e-mail: gogo5911@naver.com

G. Park

Department of Media IT Engineering the Graduate School, Seoul National University of Science
and Technology, 232, Gongneung-ro, Nowon-gu, Seoul, Korea
e-mail: gmpark@seoultech.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_23

1 Introduction

Projection mapping is a technology for visually augmenting information by projecting the virtual information onto the surfaces of the real objects that exist in the three-dimensional space. In general, projection mapping allows a number of users to simultaneously experience the augmented world without wearing the special devices and holding the equipment in their hands, which makes the immersion to be improved [1, 2].

Recently, human body projection mapping has been actively researched for artistic stage performances or education [2–6]. Human body projection mapping is a type of dynamic projection mapping technology which tracks the motion of the human and projects images on the body of the moving human. However, most methods allow the performer to move only within a limited narrow area, or to move their face or limbs in the same place [2–5]. Sometimes, the special types of equipment or sensors are attached to the human body to fastly track the body movement [6].

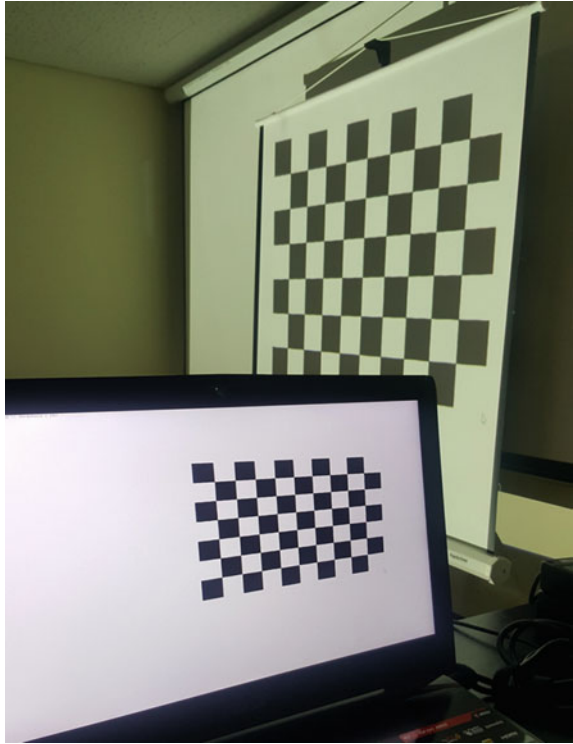
In this paper, we propose a novel dynamic projection mapping which tracks a silhouette of a performer moving freely on the stage and projects a video image on the tracked silhouette region without attaching the special sensors to the human body. In the camera-projector calibration step of the proposed method, the homography matrices for multiple depth levels are computed without the printed chessboard and the scale factors are defined for each depth level. In the dynamic projection mapping step, the silhouette of a performer is extracted based on the depth images of Kinect V2 and the orthogonal distance from the camera to the performer is computed. The silhouette image and projected video image are converted from the depth image coordinate to the window coordinate using the homography matrix of the proper depth level and the scale factors selected based on the orthographic distance. Finally, the video image is projected on the stage performer's body by drawing the video image in the converted screen region using the extracted silhouette image as a mask. In the experiments, the video images are projected onto the body of a performer moving freely, not only from left/right to right/left but also from the front/back to the back/front of the stage.

2 Camera-Projector Calibration

The camera-projector calibration is performed by applying the method of [7] in which the homography between the screen image and the camera image is computed by projecting the chessboard image on the whiteboard and capturing the projected images using a camera. Figure 1 shows the screen image and the chessboard projected on the whiteboard.

The camera image coordinates $[x, y]$ of fifty-four corner points of the chessboard are extracted using `cv::findChessboardCorners()` function of OpenCV library. In order to define the 3D stage space, the chessboard is projected on the whiteboard

Fig. 1 The chessboard image on the screen window and the chessboard which is projected on the whiteboard



located in several different depth positions which mean different distances from the camera to the whiteboard. The homography for each depth level is defined. That is, n homographies are defined for n depth levels. And the orthogonal distance from the camera to the whiteboard is also saved for each depth level. Figure 2 shows the chessboard projected on the whiteboard located in the different depth level.

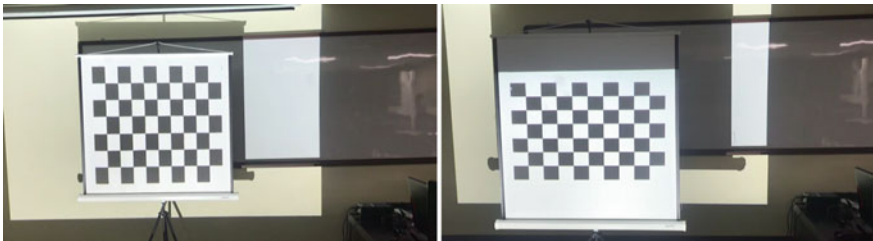


Fig. 2 The chessboard projected on the whiteboard located in the different depth level

3 Silhouette Tracking and Projection Mapping

Using Kinect V2 which includes an RGB color camera with 1920×1080 , a depth sensor with 512×424 and an infrared sensor with 512×424 , we can track 25 human body joints and silhouette. The coordinates of the body tracking points of Kinect V2 are aligned with the depth image. Since the homography computed in the camera-projector calibration step is based on the color image, the coordinates of the body points of Kinect V2 should be realigned with the color image. Figure 3 depicts the RGB color camera image and the body silhouette image realigned with the color image. The red points in Fig. 3 are the human body joint points tracked by Kinect V2.

The spine-mid point and head point among the human body joint points are selected in order to define the square region in which the video image is rendered. The length of five times the distance between the spine-mid point and the head point is set to the length of one side of the square on which the video image is drawn. The center of the square is set to the spine-mid point. The yellow square in Fig. 4 depicts the region which the video image is rendered in the RGB color camera image space. The orthogonal distance d from the camera to the spine-mid point is computed and it is used as the depth level for selecting the proper homography H^d . The yellow square and white silhouette points of Fig. 4 are aligned with the screen window space using the homography H^d . Finally, the video image is rendered in the converted yellow square by using the converted silhouette image as a mask. Figure 5 shows the masking result which is rendered in the window space.

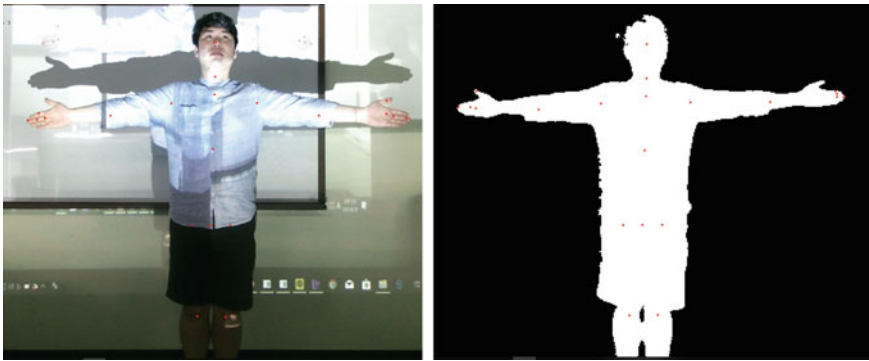


Fig. 3 The RGB color camera image (left) and the body silhouette image realigned with the color image

Fig. 4 The square region which the video image is rendered in the RGB color camera image space

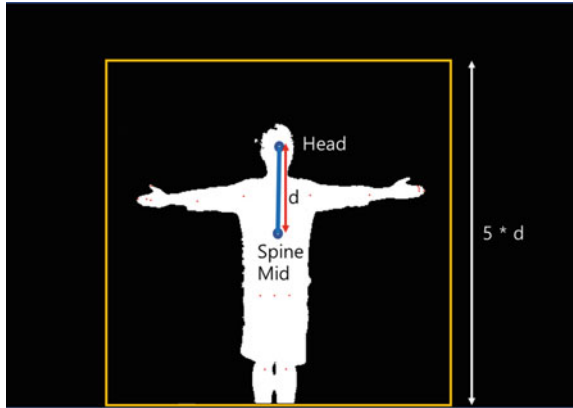
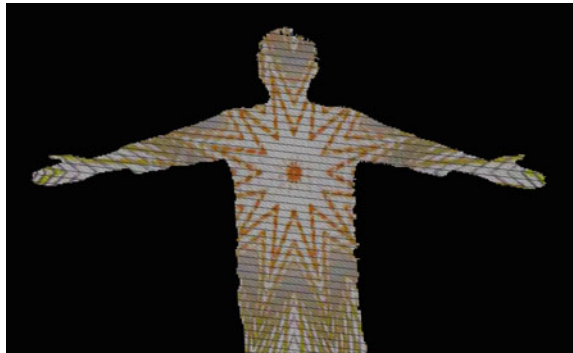


Fig. 5 The masking result which is rendered in the window space



4 Experiment Results

For the experiment, we installed a Kinect V2 and a wide-angle projector in the front of the moving whiteboard as shown in Fig. 6. Figure 7 shows the human joint points and silhouette image extracted using Kinect V2 when a performer wears large sleeved clothing. Even though most of the human joints were extracted incorrectly due to large sleeves, the head and spine-mid points were extracted correctly so that the video was rendered in the correct position while masking by the silhouette image as shown in Fig. 8. Figure 9 shows the results of the projection mapping onto the human body located in the different depth position. The spine-mid point is projected onto the middle of the body correctly, but the some errors in the silhouette boundary were shown. Therefore, it is found that the fine-tuning of the scaling is necessary after applying the homography.

Figure 10 shows the results of the projection mapping onto the panel when the performer is moving while holding the panel in his hand. The video image was projected properly onto the panel regardless of the depth level.



Fig. 6 The Kinect V2 and the projector used in the experiments

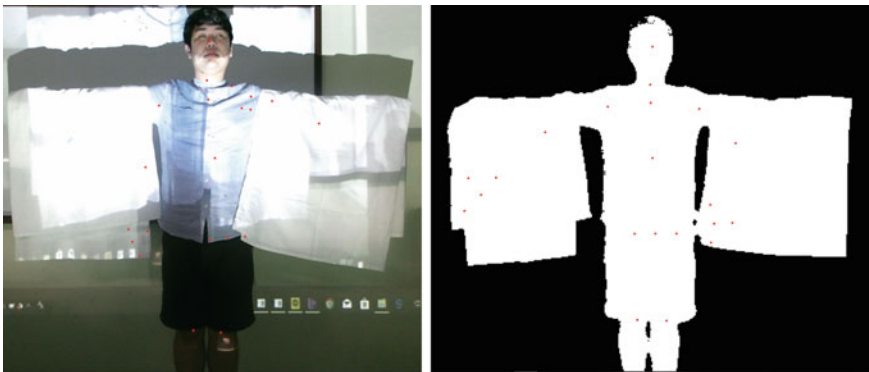


Fig. 7 The human joint points and silhouette image extracted using Kinect V2 when a performer wears large sleeved clothing

5 Conclusion

In this paper, we proposed a dynamic human body projection mapping method that allows the performer to move freely on the stage. The proposed method applied the simple camera-projector calibration that considers the different distance from the camera and doesn't require the printed calibration board. In the proposed method, the human body projection mapping is implemented by deciding the projection region based on the head and spine-mid points of the human body, and by masking using

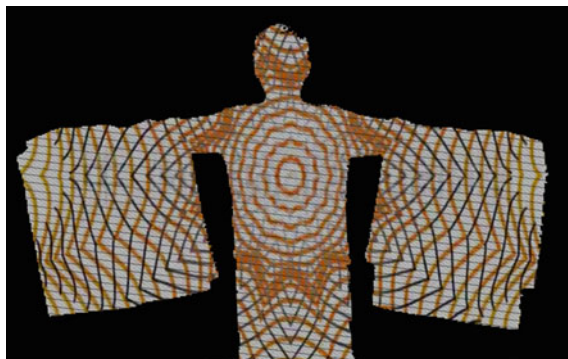


Fig. 8 The result of video rendering with masking using the silhouette image



Fig. 9 The results of projection mapping onto the human body in the different depth level

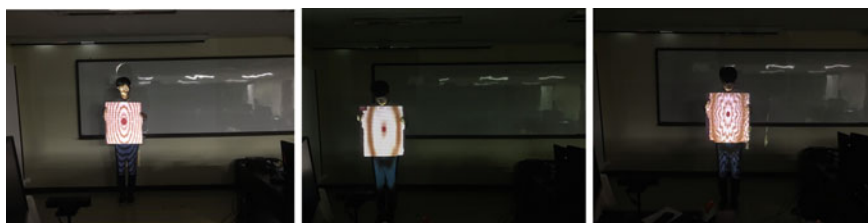


Fig. 10 The results of projection mapping onto the moving panel

the human silhouette image. As future work, we'd like to improve the accuracy of the projection mapping and to reduce the projection delay.

Acknowledgements This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF-2017R1D1A1B03035718) funded by the Ministry of Education. The human images of this paper were used under publishing consent.

References

1. Narita G, Watanabe Y, Ishikawa M (2017) Dynamic projection mapping onto deforming non-rigid surface using deformable dot cluster marker. *IEEE Trans Vis Comput Graph* 23(3):1235–1248
2. Lee J, Kim Y, Kim D (2014) Real-time projection mapping on flexible dynamic objects. In: *HCI Korea conference*, pp 187–190
3. Hoang T, Reinoso M, Joukhadar Z, Vetere F, Kelly D (2017) Augmented studio: projection mapping on moving body for physiotherapy education. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp 1419–1430
4. Interactive body projection mapping for hypermetrop. <https://vimeo.com/34609484>
5. Siegl C, Lange V, Stamminger M, Bauer F, Thies J (2017) FaceForce: markerless non-rigid face multi-projection mapping. *IEEE Trans Vis Comput Graph* 23(11):2440–2446
6. Real time tracking and projection mapping. https://www.youtube.com/watch?v=XkXr-LZmnQ_M
7. Kim S, Choi Y (2018) Dynamic projection mapping using kinect-based skeleton tracking. In: *Proceedings of the 10th International conference on computer science and its applications*

Reinforcement Learning for Rate Adaptation in CSMA/CA Wireless Networks



Soohyun Cho

Abstract Herein, we propose a reinforcement learning agent to control the data transmission rates of nodes in CSMA/CA-based wireless networks. We designed a reinforcement learning agent based on Q-learning. The agent learns the environment using the timeout events of packets, which are readily available in the data sending nodes. The agent controls the data transmission rate by adjusting the modulation and coding scheme (MCS) levels of the packets to effectively utilize the available bandwidth in dynamically changing channel conditions. We used the ns3-gym framework to simulate reinforcement learning. Simulation results show that the proposed reinforcement learning agent adequately adjusts the MCS levels according to changes in the environment and achieves a high throughput comparable to that of Minstrel, a well-known data transmission rate adaptation scheme.

Keywords Reinforcement learning · CSMA/CA · Q-learning · ns-3 · ns3-gym

1 Introduction

Currently, technologies for artificial intelligence (AI) are being developed rapidly, and they are starting to be applied to many areas. Among the many technologies for AI, we are particularly interested in reinforcement learning (RL) [1] because it can manage dynamic or interactive systems. RL uses observations and rewards from its environment in a series of time steps and applies actions to the environment for the next time steps to achieve its final goal. Q-learning [2] is a basic scheme of RL that uses a Q-table, which represents the states of the environment and the possible actions for each state. An RL agent based on Q-learning selects the best action based on the current state and the Q-values of the actions; subsequently, it applies the selected action in the next time step.

Recently, the adoption of the deep neural network [3] in RL resulted in deep RL, e.g., the deep Q-network (DQN) [4]. It has been reported that the DQN could

S. Cho (✉)

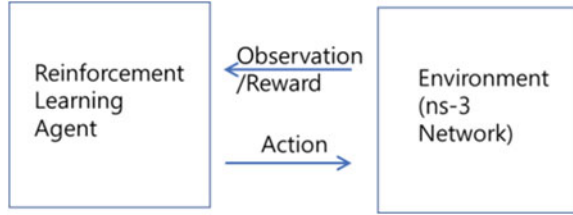
Hongik University, 94 Wausan-ro, Mapo-gu, Seoul 04066, South Korea
e-mail: cho.soohyun@hongik.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_24

175

Fig. 1 The ns3-gym framework with the ns-3 network simulator



outperform experts in some interactive games such as the Atari 2600 games [5]. However, in this study, we used the basic Q-learning scheme to control the data transmission rates of nodes in carrier sensing multiple access with collision avoidance (CSMA/CA)-based wireless networks because of the simplicity of the scheme.

In many RL studies, researchers use a simulation framework called OpenAI Gym [6], which provides a simulation environment for various applications such as the Atari games. Recently, the authors of [7] introduced a framework called ns3-gym, which provides a simulation environment resembling that of OpenAI Gym to study RL in network research. ns3-gym functions in parallel with ns-3 [8], which is a widely used network simulator. Using ns3-gym, an RL agent can learn the environment through the observations and rewards from the simulated network, and the actions selected by the RL agent for the next time steps can be applied to the simulated network, as shown in Fig. 1.

Herein, we propose a novel RL agent based on Q-learning to control the data transmission rates of nodes in CSMA/CA-based wireless networks and evaluate its performance using the ns3-gym framework.

2 Rate Adaptation in CSMA/CA Networks

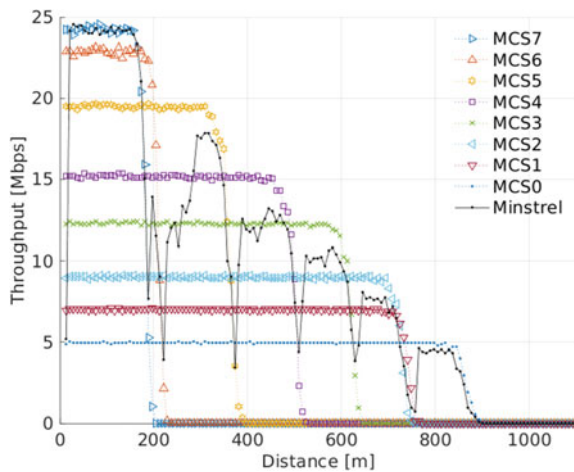
To control the data transmission rates of the nodes in CSMA/CA-based wireless networks such as the IEEE 802.11 standards [9], the nodes can select a different modulation and coding scheme (MCS) for their packets. Table 1 shows the MCSs and their physical data rates in IEEE 802.11a [10], which we used for the ns-3 simulations in this study. Each of the MCS levels requires a certain signal-to-interference and noise ratio (SINR) for the successful reception of the packets of the MCS level at their receivers.

Because a high MCS level requires a high SINR, the sender nodes must select adequate MCS levels for the data packets to their receivers to achieve a high throughput, especially when the channel conditions are changing. To accurately adjust the MCS levels, many algorithms have been proposed, such as Minstrel [11], which is a popular data transmission rate control algorithm for CSMA/CA wireless networks. Minstrel adjusts the MCS levels of data packets using the results from probing data packets at different MCS levels. Figure 2 shows the achieved applica-

Table 1 Modulation and coding schemes of IEEE 802.11a [10]

MCS level	MCS	Data rate [Mbps]
0	BPSK, 1/2 coding	6
1	BPSK, 3/4 coding	9
2	QPSK, 1/2 coding	12
3	QPSK, 3/4 coding	18
4	16-QAM, 1/2 coding	24
5	16-QAM, 3/4 coding	36
6	64-QAM, 2/3 coding	48
7	64-QAM, 3/4 coding	54

Fig. 2 Changes in application throughputs achieved by the receiver when the sender uses different MCS levels or Minstrel



tion throughputs by a receiver when its sender uses different MCS levels or Minstrel while the receiver is moving away from the sender.

The simulation results show that when a fixed MCS level is used, the sender does not fully utilize the available bandwidth of the network. With high MCS levels (such as 7 and 6), the receiver achieves a high throughput¹; however, it could not achieve any throughput beyond 200 m. With low MCS levels (such as 1 and 0), the receiver could receive packets when it was far from the sender (up to approximately 900 m for MCS level 0). However, it could not achieve a high throughput when the sender and receiver were close to each other. By contrast, from the simulation results of Minstrel, it is clear that the receiver achieves a high throughput when it is close to the sender, and that it still receives packets when the distance is approximately 900 m.

For the simulations, we used the IEEE 802.11a implementation in the ns-3 version 3.29. In the simulated network, only one sender node and one receiver node exist.

¹The achieved throughputs were less than the physical data rates in Table 1 because we measured the application throughput, and the CSMA/CA mechanism consumed bandwidth.

Both the sender and receiver operate in the distributed coordination function (DCF) [9] mode. During the simulations, the sender node sends a constant bit rate traffic of 60 Mbps to the receiver node for 15 s, while the receiver node moves away from the sender at a speed of 80 m/s. Both the sender and the receiver use 100 mW as the transmit power. For signal propagation, we used the `TwoRayGroundPropagation` model in ns-3, and fading was not considered. The noise figure was set to 7 dB. To compute the success probability of the received packet, the `NistNetErrorRate` model in ns-3 was used. The data packet size was set to 1,000 bytes, and the request-to-send/clear-to-send scheme was not used. Other settings were left unchanged from the ns-3 distribution unless otherwise mentioned.

3 Rate Adaptation Using RL

In this section, we describe the RL agent proposed to control the data transmission rates of nodes in CSMA/CA wireless networks. The ns3-gym framework that we used for RL research could report all the network situations such as the queue sizes of the nodes at each time step in the form of the observation. However, we used only the information readily available in the sender node of the CSMA/CA wireless networks, such as the contention window (CW) size. This is because we assume that each node adopts its own RL agent that operates with limited information available in the node, and we consider that a CSMA/CA wireless network in the DCF mode is inherently a distributed system of multiple participants.

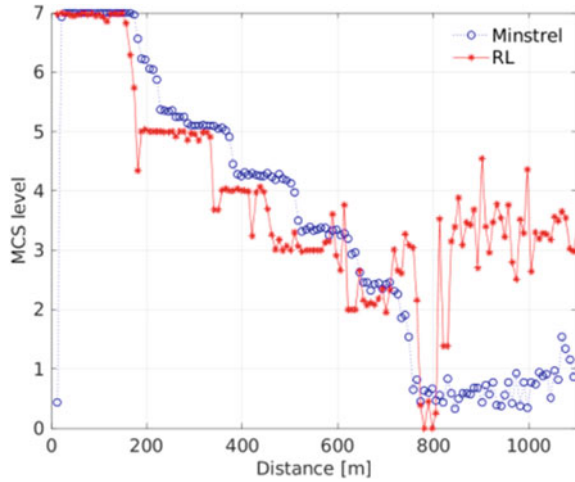
In the DCF mode, each CSMA/CA node selects its random backoff time based on the current CW size to avoid collisions with packets sent by other nodes in the network. When a CSMA/CA node sends a packet for the first time, it sets the CW size to its minimum value (CW_{\min}), which is 15 in IEEE 802.11a. Whenever the acknowledgement frame from the receiver of the packet does not arrive in a given time (i.e., a timeout event occurs), the sender doubles its CW size before the retry limit is reached (seven in the short retry limit case). Therefore, with the n th timeout event, the CW size in a node is given as Eq. 1, and the CW size can be from 15 to 1,023 (as in IEEE 802.11a) with the short retry limit.

$$CW = 2^n (CW_{\min} + 1) - 1. \quad (1)$$

We considered the CW size as an indicator of the network situation that can be used for the RL agent of the sender. However, we used the number of consecutive timeout events (i.e., n) to reduce the size of states of the Q-table in the RL agent. The number of consecutive timeout events can be obtained from the CW size using Eq. 2, and it can be an integer from 0 to 6.

$$n = \log_2(CW + 1) - 4. \quad (2)$$

Fig. 3 Changes in average MCS levels used by the sender node



We used n as the observation that an RL agent obtains in the ns3-gym framework for its sender node at each time step.

For the reward, we let the sender node count the number of acknowledged packets from the receiver during each time step and use it as the reward in the time step. The number of acknowledged packets represents the number of packets successfully received by the receiver. With the given observation and the reward for a time step, the RL agent selects the best MCS level from the Q-table, and the ns3-gym framework applies it to the sender node as the action that will be used by the sender node for the next time step. The time step was set to 1 ms considering the simulated network topology and the network parameters, such as the frequency (5.18 GHz) and the bandwidth (20 MHz) used for the simulations.

4 Simulation Results

To evaluate the performance of the proposed RL agent in Sect. 3, we performed ns-3 simulations in the ns3-gym framework with the same scenario used for the fixed MCSs and Minstrel in Sect. 2 (i.e., the receiver is moving away from the sender). We performed 10 episodes with different seed values for randomness. Figure 3 shows changes in the average MCS levels used by the sender for every 0.1 s in the 10th episode² along with those used by Minstrel. The simulation results show that the RL agent adequately adjusts the MCS levels according to the changes in distance between the sender and receiver. Furthermore, the simulation results show that the RL agent adjusts the MCS levels more accurately than Minstrel except when the receiver moves beyond 800 m.

²After several episodes for learning, the following episodes showed similar results.

Fig. 4 Changes in application throughput achieved by the receiver

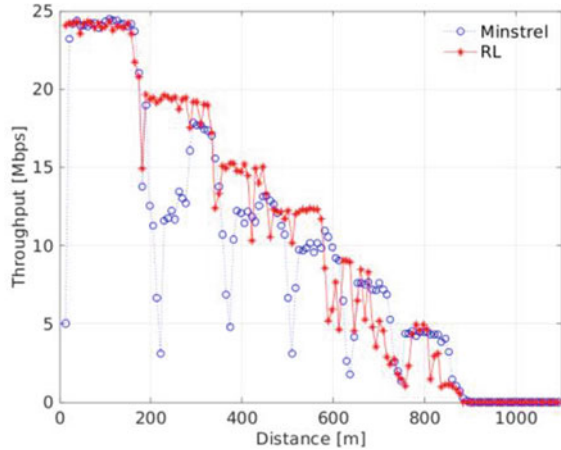


Figure 4 shows the changes in the throughput achieved by the receiver in the 10th episode for every 0.1 s and also that achieved by Minstrel. The simulation results show that the sender node with the RL agent can utilize the available bandwidth while the receiver is moving away from the sender. By comparing the results with those of Minstrel shown in Fig. 4, it is clear that the RL agent can achieve a performance comparable to that of Minstrel, which uses probing data packets.

5 Conclusions

Herein, we introduced an RL agent based on Q-learning to control the data transmission rates of CSMA/CA nodes. The RL agent learned the environment using the readily available information in the nodes and controlled the MCS levels of the data packets. Simulation results using the ns3-gym framework indicated that the proposed RL agent adequately adjusted the MCS levels of the data packets sent to the receiver after several episodes for learning. Consequently, a sender node with the RL agent could effectively utilize the available bandwidth while the receiver node of the packets moved away. Comparing the simulation results with those of Minstrel, we demonstrated that the RL agent could achieve a performance comparable to that of Minstrel.

In future studies, we will consider adopting a deep neural network for the RL agent and perform experiments in more complex scenarios.

Acknowledgements This work was supported by 2018 Hongik University Research Fund.

References

1. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction, 2nd edn. Bradford Book, Cambridge
2. Watkins CJ, Dayan P (1992) Q-learning. *Mach Learn* 8:279–292
3. Bengio Y (2009) Learning deep architectures for AI. *Found Trends Mach Learn* 2:1–127
4. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK et al (2015) Human-level control through deep reinforcement learning. *Nature* 518:529–533
5. Bellemare MG, Naddaf Y, Veness J, Bowling M (2013) The arcade learning environment: an evaluation platform for general agents. *J Artif Intell Res* 47:253–279
6. OpenAI Gym. <https://gym.openai.com>
7. Gawlowicz P, Zubow A (2019) ns-3 meets OpenAI gym: the playground for machine learning in networking research. In: ACM international conference on modeling, analysis and simulation of wireless and mobile systems. Miami Beach, USA
8. ns-3: a discrete-event network simulator for Internet systems network simulator. <https://nsnam.org>
9. IEEE Std. (2016) 802.11-2016 Part 11: wireless lan medium access control (MAC) and physical layer (PHY) specifications. IEEE Comput Soc
10. IEEE Std. (2003) 802.11a-1999(R2003) Part 11: wireless lan medium access control (MAC) and physical layer (PHY). IEEE Comput Soc
11. Rate Adaptation for 802.11 Wireless Networks: Minstrel (2010). <https://blog.cerowrt.org/papers/minstrel-sigcomm-final.pdf>

Biometric-Based Seed Extraction Scheme for Multi-quadratic-Based Post-quantum Computing



Aeyoung Kim and Seung-Hyun Seo

Abstract PQC (Post-Quantum Cryptography) is rapidly developing in order to provide stable and reliable quantum-resistant cryptography throughout the industry. Like the existing public key cryptography, it does not prevent a third-party using the secret key when the third party obtains the secret key by deception, unauthorized sharing, or unauthorized proxying. The most effective alternative to prevent such an illegal use is the utilization of biometrics. In this paper, we propose a biometric-based seed extraction scheme by applying PCA (Principal Component Analysis)-based CI (Confidence Interval) as a feature extraction method. The bio-seed can be a good helper data to generate biometric-based secret key in PQC such as Rainbow. It is helpful to prevent using the secret key by an unauthorized third party through biometric recognition as well as to generate a shorter secret key.

Keywords Biometric-based seed · Seed extraction · Key generation · Biometric cryptography · Post-quantum cryptography

1 Introduction

The high speed PQC (Post-Quantum Cryptography) is the latest public key cryptographic technique that has been actively studied for the emergence of quantum computers, which threaten the security of the industry's existing cryptosystems. The

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1A6A3A01013588) and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2018-0-01417) supervised by the IITP (Institute for Information & Communications Technology Promotion).

A. Kim · S.-H. Seo (✉)
Hanyang UniversityERICA, 55 Hanyangdaehak-ro, Ansan, Gyeonggi-do 15588, Korea
e-mail: seosh77@hanyang.ac.kr

A. Kim
e-mail: aeyoung@hanyang.ac.kr

existing public key cryptographic algorithms such as RSA, ECDSA, etc. are no longer secure, because all of such cryptographic algorithms are based on the difficulty of factorization or discrete logarithm problems that can be decrypted by the quantum computers. Therefore, the need for a new public key cryptography technique resistant to the quantum computation of quantum computers goes up, and quantum resistant cryptographic algorithms are being actively researched.

The basic building blocks for designing quantum resistant cryptographic algorithms are lattice, code, hash, and MQ (multivariable quadratic). Among them, the MQ signature algorithm shows faster performance than that of the existing RSA and ECDSA [1]. The representative MQ signature algorithm, rainbow signature scheme, with Quantum Security Level 128-bit is 145 times faster than RSA's and 3 times faster than ECDSA's for signature generation, and 2 times faster and 7 times faster, respectively, for signature verification.

However, like the existing encryption algorithm, the quantum-resistant cryptographic algorithm does not block the use of the third party's encryption key by deception, unauthorized sharing or unauthorized proxying of the key. The most effective alternative to prevent unauthorized third-party use of cryptographic key-based security solutions such as digital signatures and encryption algorithm is the utilization of biometrics on the existing crypto system.

Biometrics such as face can be used for user authentication because it is a unique characteristic that cannot be separated from the original owner. However, biometrics is unstable information obtained with different values with noise. This is true even though the same part of the same person is repeatedly acquired many times at short intervals such as seconds with the same device under the same conditions. In the conventional cryptosystem, since input values having a difference of one-bit lead to a very large difference for authentication, the most important problem for generating the key from the biometrics is how to find a scheme which can stably reproduce the exact bit string that looks random using the noisy input.

Some researchers have studied how to generate or bind a biometric-base key and cryptography. Their biometric-based key generation or binding schemes have been attracting much attention for additional authentication. However, biometric-based helper data extractors in their schemes are still difficult to realize due to noise and error elimination problems, computation problems, and storage size problems for helper data.

In this paper, we propose a biometric-based seed extraction scheme for using helper data or key in a post-quantum cryptographic algorithm that can effectively apply biometrics and satisfy a certain level of security such as quantum security level in 128-bit. Our target post-quantum cryptographic algorithm is Rainbow signature scheme [2], which is a kind of MQ-based public key cryptography. This scheme can also be quantum-resistant and implemented in resource-constrained IoT devices, performing at high speeds. The proposed scheme gives us usable biometric-based data as a seed to be a key which is smaller than the key in Rainbow and identifies an authorized user.

2 Related Works

2.1 *Biometric-Based Helper Data Extraction Schemes*

Biometric cryptosystem applied to retrieve or generate keys from biometrics. Because biometric variance makes it difficult to extract consistent keys directly, many related researches have addressed by “fuzzy” concepts and helper data, which is driven from biometrics: fuzzy commitment scheme, fuzzy extraction or secure sketch scheme, and fuzzy vault scheme. These approaches focus on how these systems deal with biometric variance. For this, some schemes apply error correction codes and others introduce filter functions, correlations, and quantization. While these method in binding helper data with keys are independent on biometrics, they for generating key from helper data are dependent upon biometrics, i.e. biometric features take influence on the constitution of keys.

The quantization of biometric features can be used to construct helper data or to directly generate keys or hash values. In order to provide keys, most schemes provide a parameterized encoding of intervals. Generally, in the quantization-based scheme, appropriate feature intervals to quantize are defined for each single biometric feature based on its variance. In this approach, Davida et al. [3] proposed the private template scheme, which serves a secret key and error correction check bits as a helper data. Feng and Wah [4] proposed a private key generation scheme in 40 bits of security level with “good-quality” signature-based helper data, which is extracted by using the user boundary as an interval by defined them. Li et al. [5] proposed a fuzzy extractor, which is by using pre-aligned fingerprints and Principal component analysis. This scheme concerned about the strength of the secret key extracted by using definition of interval bound, which is the minimum distance between codewords and an upper bound based on the logarithm of the Varshamov-Gilbert Bound. Rathgeb and Uhl [6] also proposed a biometric-based key generation scheme by using interval-mapping scheme. These several approaches have been published applying Interval to fingerprints and faces. However, there are commonly the limitation of bit length related to security level of keys. It is not enough for post-quantum cryptography and does not show the feasibility as much as the fuzzy extractor does.

The fuzzy extractor proposed by Dodis et al. [7, 8], one of the representative studies for solving the problem, generates a key from biometrics so that it can be applied to any public key cryptographic algorithm. However, the proposed fuzzy extractor has high computational complexity for getting helper data and a very large space in units of Giga or Tera to store the helper data. To solve this problem, Fuller et al. proposed a model using the LWE (learning with error) problem. Cheon et al. [9] also applied the LWE problem to the model proposed by Canetti et al. [10] to solve the problem and create models with more realistic storage sizes.

Unlike a general biometric-based key generation model that can be applied to any cryptographic algorithms, Conti et al. [11] proposed Biometric RSA, which is a model for generating private keys from two fingerprints stored on smart cards by mapping a prepared large prime number list, for targeting RSA. Thus, models for

generating a secret key of a specific cryptographic algorithm like RSA have been proposed, but most of them are simple mapping or linking methods like Biometric RSA. It is difficult to find a representative study because their proposals do not provide a reasonable and concrete method. Some researchers tried to use other noise sources such as multifactor, PUF (Physical Unclonable Function), and image link, but the result is also insignificant.

2.2 PCA-Based Feature Extraction

PCA (Principal Component Analysis), one of the most popular multivariate statistical techniques that uses an orthogonal transformation, has been broadly applied to multivariate data analysis, pattern recognition, and signal processing [12, 13]. It has also driven variants of PCA such as Quaternion PCA, L1-norm PCA, patch-based PCA, and $2D^2$ PCA in various aspects [14]. Such PCA has been widely used in biometrics. Especially, PCA applied to face images is called Eigenfaces, which were first developed by Sirovich and Kirby for recognition and used by Turk and Pentland [15] in face classification. They tried to find a lower-dimensional space for the simplest approach to recognize faces as a template matching problem.

3 Biometric-Based Seed Extractor

The proposed scheme extracts a bio-seed, called BS , from noisy data such as biometrics to apply MPKC. This BS is used as a secret key for the generation of a private key in MPKC. In order to design the BS extraction scheme, we consider some properties of MPKC such as Rainbow as follows:

- The MQ signature scheme can be applied over galois field $GF(256)$ as well as many other signals including images that can be represented in 8 bits.
- The private key consists of the central map Q and two affine transformations S^{-1} and T^{-1} . These are made of random numbers over $GF(256)$ as well as acquired biometrics, which also appear as random numbers because of their variance noise.
- The central map is a subspace of $n \times n \times m$ matrix, and this appears as a subspace of K^3 matrix simply in terms of size, where $K \geq n > m$.

The basic idea of the proposed scheme is to generate biometric-based secret information BS from F^{NNM} to F^{3K} by using signal processing. The BS can be used to derive a map from F^{3K} to F^{NNM} and to replace the existing central map with the BS -based central map in the private key in MQ-based PQC. The basic processes of this scheme include 4 stages as shown in Fig. 1. The input of the first to third stages is deleted after it is used. BS is the only information that must be kept in a secure area as SK , as it is impossible to be regenerated without using the same biometrics. BS is stable data for the mid-term. Our chosen methods (Cm) for each stage are camera

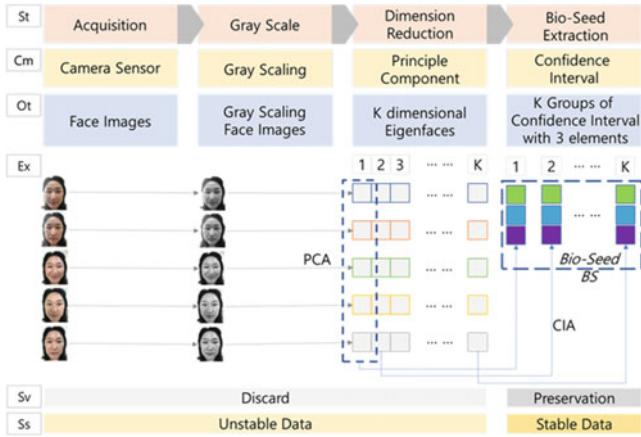


Fig. 1 The proposed bio-seed extraction process (*St*: step name, *Cm*: chosen method, *Ot*: output type, *Ex*: example, *Sv*: discard or save mode, *Ss*: state of data stability)

for acquisition, gray scaling for transforming into the grayscale, PCA for dimension reduction, and CIA (Confidence Interval Analysis) for *BS* extraction.

- (1) *Acquisition Stage*: The $N \times N$ face images $I = [I_1, \dots, I_M]$ are acquired.
- (2) *Gray-scaling Stage*: The acquired images are converted into grayscale image $GI = \{GI_1, \dots, GI_M\}$. GI is represented by 8 bits, so that each pixel is the same as an unsigned byte in various computer programming languages such as C. Each pixel has 256 difference values $\{0, \dots, 255\}$ and can be applied for operations over $GF(256)$.
- (3) *Dimension Reduction Stage*: We used PCA for dimension reduction of the grayscale face image GI by projecting it from an N^2 -dimensional space to a K -dimensional space, where $K \gg N^2$ and $K \geq n$. In this paper, the PQC model used to fuse the face images is Rainbow, and the central map is replaced with a *BS*-based central map. Therefore, we need n dimensions at least, where n is a parameter in Rainbow.
- (4) *Bio-Seed Extraction Stage*: We consider that the $n \times n \times m$ central map can be a subspace of a K^3 bigger cube as a candidate of the new central map. To use this for generating the cube in the key generation stage, we represent the dimension reduced faces into K groups including upper bound, median, and lower bound by CIA, and call them *BS* as *SK* for generating a private key map and a public key map of Rainbow (Table 1).

The target post-quantum cryptographic algorithm in this experiment is Rainbow to show how the proposed bio-seed extractor can be applied to MQ signature schemes. The parameter set $(F_2^8, 36, 21, 22)$ for Rainbow is considered the optimal secret key size at a 128-bit post-quantum security level. The parameters for RSA and ECDSA are 3072 and 256^e are also selected at 128-bit post-quantum security level for a comparison. When the PQ Security Lv. is 128 bits, *SK* is 237 bytes and about 92.3

Table 1 Size (byte) of signature, PK, and SK

Schemes	QSLv	Signature	PK	SK
RSA	128	361	384	3,072
ECDsa	128	64	64	96
Rainbow	128	79	139,320	105,006
Bio-seed-based rainbow	128	79	139,320	237

and 99.8% shorter than that of RSA and Rainbow. The total key size is 139,557 bytes which is about 42.9% shorter than that of Rainbow.

4 Conclusion

We proposed a new biometric-based seed extraction scheme for a multivariate quadratic-based signature scheme such as Rainbow, which is one of PQC algorithm. When the proposed scheme was applied in rainbow, we can get the suitable bio-seed as a helper data, which can be applied to generate or bind with the secret key in key pair, the public key in key pair, the vinegar variables in the first layer. To the best of our knowledge, such a fusion model is being proposed for the first time.

Since the bio-seed can be get by biometric recognition in the proposed scheme, it prevents an unauthorized user using the secret key. When we apply PCA-based CIA as the feature extraction method for the bio-seed extraction, the secret key by using the bio-seed is shorter than the length of the secret key in Rainbow. We believe that this proposed scheme is a practical biometric-based seed extractor to generate the secret key for quantum-resistant cryptography. We also expect that this belief will be proven in the future by experimenting with more feature extraction methods, various biometric templates, and biometric sensors.

References

1. Yi H (2018) Under quantum computer attack: is rainbow a replacement of RSA and elliptic curves on hardware? *Secur Commun Netw*
2. Ding J, Schmidt D (2005) Rainbow, a new multivariable polynomial signature scheme. In: International conference on applied cryptography and network security. Springer, Heidelberg, pp 164–175
3. Davida GI, Frankel Y, Matt B (1998) On enabling secure applications through offline biometric identification. In: IEEE international symposium on security and privacy. IEEE Press, pp 148–157
4. Feng H, Wah CC (2002) Private key generation from on-line handwritten signatures. *Inf Manag Comput Secur* 10:159–164

5. Li Q, Guo M, Chang EC (2008) Fuzzy extractors for asymmetric biometric representations. In: 2008 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE Press, pp 1–6
6. Rathgeb C, Uhl A (2009) An iris-based interval-mapping scheme for biometric key generation. In: 6th international symposium on image and signal processing and analysis. IEEE Press, pp 511–516
7. Dodis Y, Ostrovsky R, Reyzin L, Smith A (2008) Fuzzy extractors: how to generate strong keys from biometrics and other noisy data. *SIAM J Compt* 38:97–139
8. Verbitskiy EA, Tuyls P, Obi C, Schoenmakers B, Skorie B (2010) Key extraction from general nondiscrete signals. *IEEE Trans Inf For Sec* 5:269–279
9. Cheon JH, Han K, Kim J, Lee C, Son Y (2016) A practical post-quantum public-key cryptosystem based on spLWE. In: International conference on information security and cryptology. Springer, Cham, pp 51–74
10. Canetti R, Fuller B, Paneth O, Reyzin L, Smith A (2016) Reusable fuzzy extractors for low-entropy distributions. In: Annual international conference on the theory and applications of cryptographic techniques. Springer, Heidelberg, pp 117–146
11. Conti V, Vitabile S, Sorbello F (2012) Fingerprint traits and RSA algorithm fusion technique. In: 6th international conference on complex, intelligent, and software intensive systems. IEEE Press, pp 351–356
12. Bouwmans T, Javed S, Zhang H, Lin Z, Otazo R (2018) On the applications of robust PCA in image and video processing. In: The IEEE. IEEE Press, pp 1427–1457
13. Ameer B, Belahcene M, Masmoudi S, Hamida AB (2019) Hybrid descriptors and Weighted PCA-EFMNet for Face Verification in the Wild. *Int J Multimedia Info Retriev* 8(3):143–154
14. Hamill JM, Zhao XT, Mészáros G, Bryce MR, Arenz M (2018) Fast data sorting with modified principal component analysis to distinguish unique single molecular break junction trajectories. *Phys Rev Lett* 120
15. Turk M, Pentland A (1991) Eigenfaces for recognition. *J Cogn Neuro* 3:71–86

Lighting System to Maintain Color Temperature of Natural Light by Reflecting Changes of the Incoming Light



Se-Hyun Lee, Seung-Taek Oh, and Jae-Hyun Lim

Abstract This paper proposes a lighting system that maintains color temperature of natural light by reflecting the changes of external incoming light. For that, characteristics of indoor lighting environment are measured and experimental environment to control the light is constructed. The color temperature of the natural light by solar terms are derived by analyzing measured natural light database. After that, color temperatures from various points inside the room are measured and LED lighting is controlled by linking it with the measured color temperature data. The results indicated that the lighting system that adaptably reproduce the color temperature of natural light even if external incoming light is changed. The color temperature changes of the indoor lighting environment are measured and analyzed before and after implementing the proposed system to validate the performance of the proposed lighting system that maintains color temperature of natural light.

Keywords Natural light · Biorhythm · Color temperature · Lighting system

1 Introduction

Sunlight is being changed at 24 h interval and is closely related with human health. Especially, color temperature among the characteristics of sunlight is known to greatly affect the human biorhythm [1, 2]. However, as indoor activity time of modern people increases and most of the indoor lighting environment provide a uniform color temperature, a problem of imbalance of biorhythm is raised [3, 4].

S.-H. Lee (✉) · J.-H. Lim

Department of Computer Science & Engineering, Kongju National University, Cheonan, Republic of Korea

e-mail: bunkerbuster@smail.kongju.ac.kr

J.-H. Lim

e-mail: defacto@kongju.ac.kr

S.-T. Oh

Smart Natural Space Research Center, Kongju National University, Gongju-si, South Korea

e-mail: ost73@kongju.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_26

In order to resolve this problem, methods to provide color temperature and circadian characteristics of natural light have been introduced [5]. However, most of the indoor lighting environment is constructed with natural light and artificial light, therefore indoor lighting environment is changed by incoming natural light, which consequently disable maintaining intended lighting environment.

In this paper, a lighting system that maintains color temperature of natural light by reflecting changes of external incoming light is proposed. First, the characteristics of color temperature are analyzed based on the measured natural light database and the color temperature standard by solar terms is extracted to control the light. In addition, an experimental environment is constructed to investigate characteristics of indoor lighting environment, and changes of color temperature inside the room by external incoming light are measured. After that, individual control for the indoor LED lightings is realized so that the color temperature of indoor lighting environment that is changing by incoming external light would be in conformity with color temperature standard by solar terms. In addition, possibility of color temperature reproduction of natural light through implementing the proposed system is examined through experiment conducted under the environment of changed incoming light.

2 Lighting System to Maintain Color Temperature of Natural Light

The proposed lighting system was constructed to adapt and reproduce the color temperature characteristics of natural light. A lighting system composed of sensors, controllable LED lighting, and measured natural light DB was constructed. Figure 1 is a schematic diagram of the proposed lighting system.

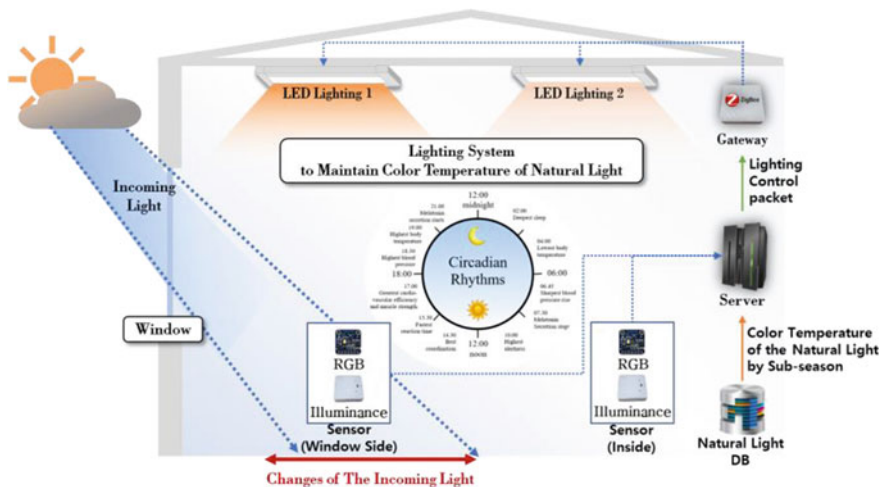


Fig. 1 Schematic diagram of lighting system to maintain color temperature of natural light

The color temperature and illuminance in Fig. 1 were measured by using RGB sensor (TCS34725) and illuminance sensor (TSL 4531). The color temperature values were used after comparing and calibrating the output values through RGB based calculation with the spectral radiometer (CAS 140CT, Instruments). In addition, the indoor lighting was constructed with Warm/Cool 2 channels and LED lighting that could realize color temperature of 2687–6000 K and illuminance of 0–1000 lux through controlling the 256 stages per each channel. In addition, ZigBee based wireless communication environment was adopted so that lighting control and sending and receiving the sensed data could be realized. Additionally, a server and a gateway were implemented to support loading the natural light data base and operation of lighting system.

The circadian characteristics of the color temperature by solar terms were analyzed based on the measured natural light database in order to draw control standard of lighting to reproduce color temperature of natural light. The spectral radiometer (CAS 140CT, Instruments) was used for analysis from March 2017 till Aug. 2019 by using measured natural light database. The circadian characteristics of color temperature of natural light by solar terms are as shown in Fig. 2. While, Table 1 presents the color temperature analysis results for the subdivision of solar terms (‘Cheoseo’) that fell nearest time to the lighting control experiment.

The proposed lighting system was adjusted its initial control stage to Warm = 0 and Cool = 0, and color temperatures and illuminances from various points in the room were measured. The illuminance and color temperature at each point that are measured in real-time were compared if these were in conformity with reference color temperature in Table 1. When difference was observed, repeated control function was

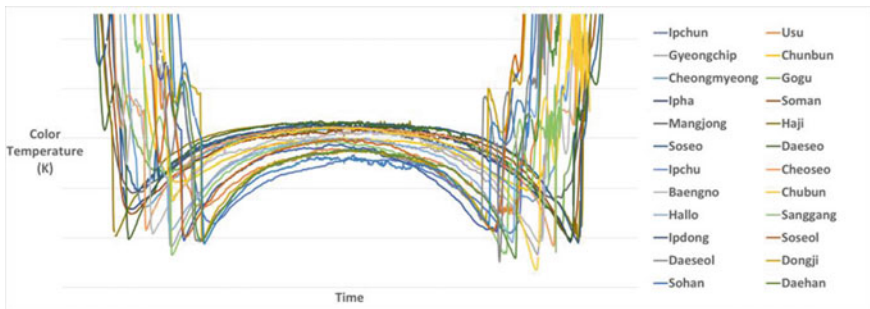


Fig. 2 Characteristics analysis for color temperature of the natural light by solar terms

Table 1 Example of characteristics analysis for color temperature of natural light by solar terms (‘Cheoseo’)

Time	06	07	08	...	12	...	16	17	18
Color temperature (K)	5565	5650	5699	...	5812	...	4778	4376	3922

realized by LED lighting stage that adjusted the color temperature at each point in the room to be near to the reference color temperature.

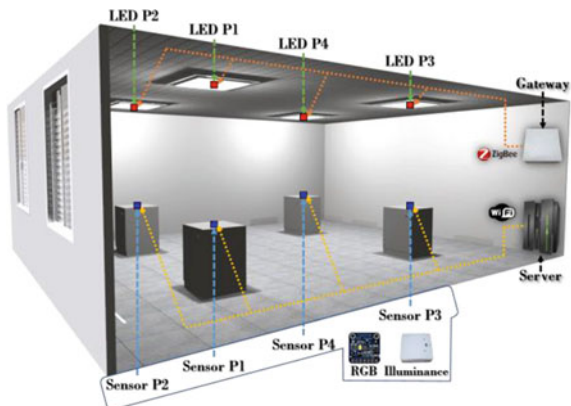
3 Experiment and Evaluation

An experimental environment was constructed to measure optical characteristics changes at each point of room after implementing the proposed system. Experimental space was set at laboratory in the 4th floor of building located at latitude of 36° and longitude of 127° . The size of experimental setup was $630 \times 720 \times 270$ cm (W \times L \times H) and two windows of dimension 140×165 cm (W \times L) were located towards south. Figure 3 shows the perspective drawing of the experimental set up.

Experiment was conducted from 06:00–19:00 on the Aug. 21, 2019. The weather conditions on the day of the experiment corresponded to an average cloudiness of 6.8 (heavily cloudy, based on data provided by the Korea Meteorological Administration). The experimental environment was categorized into the experimental group which was the proposed lighting system environment that reproduced the color temperature of natural light by adapting the changes of external incoming light and a control wherein general color temperature reproduction method was applied. The illumination in each of the experimental lighting environment was maintained to 400–500 Lux. In addition, the optimal angle of the blind that can maintain above illumination was derived as 150° in the preliminary experiment and this angle was maintained all the time during experiment. The changes of color temperature characteristics during one day under each environment were measured using RGB sensors and these were compared. Figure 4 shows experimental result for the control group (left) and the experimental results for the experimental group (right).

The color temperatures at each point (P1–P4) in the room were uneven even the environment was arranged as reproducing the color temperature of natural light as shown in the control group results as in Fig. 4. Particularly, at the window where

Fig. 3 Perspective drawing of experimental environment



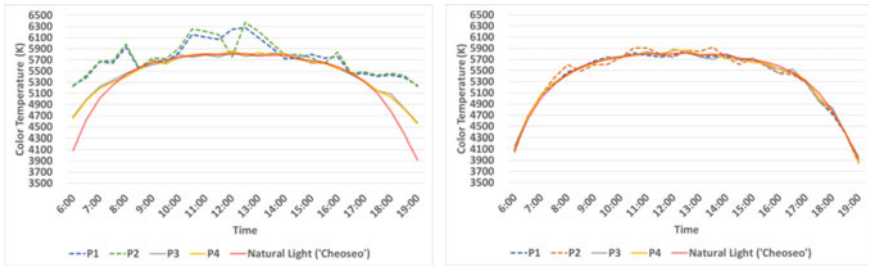


Fig. 4 Color temperatures under indoor lighting environment: control group (left) and experimental group (right)

effect of the external incoming light was greatest, the color temperature variation was severe, resulting non-conformity with the intended color temperature standard of natural light. Yet, the inner side of the room (P3, P4) did not match with the color temperature standard only in some areas before 8:00 and after 18:00. Whereas, in the experimental group results as in Fig. 4 wherein the proposed lighting system was applied, the color temperatures at each point (P1–P4) in the room were similar to that of natural light. With these experimental results, color temperature characteristics of natural light could be perfectly reproduced by adapting to the changes of external incoming light through the proposed system that measured color temperature of each point in the room and lighting was controlled.

4 Conclusion

In this paper, a lighting system that maintains color temperature of natural light by reflecting the changes of external incoming light is proposed. First, a color temperature standard by solar terms was derived by analyzing color temperature characteristics of natural light by using the natural light DB. The optical characteristics of indoor lighting environment were measured by implementing RGB and illuminance sensor and LED lighting that could be individually controlled and an experimental environment was constructed to control the lighting. Furthermore, a lighting control system was realized by implementing server that supported loading the measured natural light DB and collection of the sensed data and gateway that supported ZigBee based wireless communication. The color temperatures for each point in the indoor lighting environment were measured and a control function of the lighting that repeatedly performed LED lighting by channel to comply the color temperature standard in real-time was realized. The changes of the color temperature characteristics in the indoor lighting environment before and after implementing the proposed lighting system were measured and compared. The comparison results showed that in the control group which was the lighting environment that reproduced the color temperature of natural light, color temperature changes were severe due to incoming external light.

Particularly, the color temperatures were not in conformity with the color temperature standard in almost all the time zones at the window side (P1, P2). Whereas, with the proposed lighting, the color temperatures in all the points of window (P1, P2) and inner side (P3, P4) of the room were similar with that of natural light in all the time zones. The results confirmed that when the proposed system is implemented, the color temperature of natural light can be perfectly reproduced even under changes of external incoming light.

Researches are needed to expand the propose lighting system in the future. For the purpose, research and validation experiment would be performed to save lighting energy and to improve lighting quality through implementing a blind control.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A2B2005601).

This work was supported by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1A6A1A03032988).

References

1. Kim KS (2013) Analysis of melatonin suppression according to different color temperature in white LED light. Master thesis, Sejong University, Seoul, Republic of Korea
2. Kwon SY (2013) Design of natural light reproduction system using a matching algorithm based on chromaticity coordinates. Doctoral thesis, Kongju National University, Cheonan, Republic of Korea
3. Lee JE (2004) Development and application of health lighting plan for quality of life in residential areas. Sejong University Seoul, Republic of Korea
4. Kim IT (2011) A study of control algorithm for development of an adjustable CCT LED lighting system in architectural space. Sejong University Seoul, Republic of Korea
5. Kim KM, Kim YW, Oh ST, Lim JH (2013) Analysis of melatonin suppression according to different color temperature in white LED light. Sejong University, Seoul, Republic of Korea (2013)

Deep Neural Network Model for Calculating Ultraviolet Information with Seasonal Characteristics from Illuminance



Deog-Hyeon Ga, Dae-Hwan Park, Seung-Taek Oh, Heon-Tag Kong, and Jae-Hyun Lim

Abstract Ultraviolet rays have beneficial or detrimental effects on human health, depending on the degree of exposure. Recently, with increasing interest about health, the demand for UV-related information has also increased. Therefore, this paper proposes a DNN model that applies seasonal characteristics to calculate the illuminance-based UV information at the user's location. For that, the relationship between the UV information and natural light's characteristics using the measured natural light data was analyzed. After that, a data set was constructed by considering seasonal characteristics of the light, and DNN model learning is performed through the data set. The DNN model which calculates illuminance-based UV information is validated by considering accuracy of UVI calculation according to adaptation of seasonal characteristics in the input data. The proposed model was validated by comparing accuracy of UVI calculation according to application of the seasonal characteristics for the DNN model that determines the illuminance-based UV information.

Keywords Deep neural network (DNN) · Ultraviolet index (UVI) · Seasonal characteristics · Natural light

D.-H. Ga (✉) · D.-H. Park · H.-T. Kong · J.-H. Lim
Dept. of Computer Science & Engineering, Kongju National University, Cheonan, Republic of Korea

e-mail: 201401902@smail.kongju.ac.kr

D.-H. Park
e-mail: glow153@smail.kongju.ac.kr

H.-T. Kong
e-mail: htkong@kongju.ac.kr

J.-H. Lim
e-mail: defacto@kongju.ac.kr

S.-T. Oh
Smart Natural Space Research Center, Cheonan, Republic of Korea
e-mail: ost73@kongju.ac.kr

1 Introduction

Ultraviolet rays have beneficial effects on human health such as vitamin D synthesis, maintenance of calcium homeostasis and prevention of osteoporosis when proper exposure is achieved. However, overexposure can have a detrimental effect on skin health, including skin burns and skin cancer [1]. Recently, as people's interest in health increases, the demand for UV-related information that has a great influence on the human body is also increasing. Accordingly, various studies are being actively conducted to provide platforms and services for calculating UV information [2].

In general, users must use information services provided by the Korean Meteorological Agency or use data provided by the relevant agencies in order to obtain ultraviolet information. However, the UV information provided by the Meteorological Agency is based on the data measured at limited 15 domestic UV stations. Therefore, it does not accurately reflect the characteristics of the user's location. Furthermore, in order to obtain more accurate UV information, it is necessary to use specialized measuring equipment or to operate a separate measuring module, but there is a limit of requiring related expertise and the module itself would not satisfy user convenience.

Therefore, this paper proposes a deep neural network model on which seasonal characteristics of the UV are applied to calculate illuminance-based UV information. For that, expertise light measuring equipment (spectral radiometer) is operated. The measured and collected natural light data were used to analyze correlation between the light characteristics such as illuminance and UV index (UVI). A data set was then extracted for the learning of the deep neural network model for which the seasonal characteristics of UV are considered. After that, a deep neural network model with the input data of illuminance and seasonal characteristics was established to calculate more accurate illuminance-based UV information. In addition, the UVI calculation results of the deep neural network model before and after applying the seasonal characteristics of UV were compared to validate the performance of the proposed model.

2 Analysis of Natural Light Data and Seasonal Characteristics

Before calculating the illuminance-based UV information, the measured natural light data such as illuminance and UVI were analyzed. In addition, the seasonal characteristics of natural light were also investigated. For the analysis, the natural light data measured and collected from the roof of 10-story building of this University located at latitude of 36.85° and longitude of 127.15° over about two years (Apr. 1, 2017–Aug. 13, 2019) were used. Natural light data were collected separately in terms of light characteristics such as illuminance, color temperature, chromaticity coordinate, and UV index. The intensity of domestic ultraviolet rays is not

constant and changes year-round [3]. In order to analyze the domestic UV rays considering such characteristics, the hourly solar zenith angle information provided by the Korea Astronomical Research Institute were also collected in constructing database. Table 1 shows the correlation between the characteristics such as natural light, UVI, and solar zenith angle.

The analysis results presented in Table 1 show that the correlation of UVI with illuminance is high with the value 0.76, and the correlation with the solar zenith is also high with the value -0.93 . However, the color temperature (CCT) and color coordinate show relatively low correlation of 0.44 and -0.53 , respectively. In addition, in case of the illuminance, it showed a relatively low correlation with UVI as compared to solar zenith angle. The results were analyzed by output of scatter plots between the illuminance and UVI classified by each season of spring (March, Apr., and May), summer (June, July, and Aug.), autumn (Sept, Oct., and Nov.), and winter (Dec., Jan., and Feb.) for the natural light data. The analysis results are presented in Fig. 1.

Table 1 Correlation between natural light data

	Illuminance	CCT	Color coor-dinate x, y	Solar zenith
UVI	0.7659	0.4461	-0.5347	-0.9314

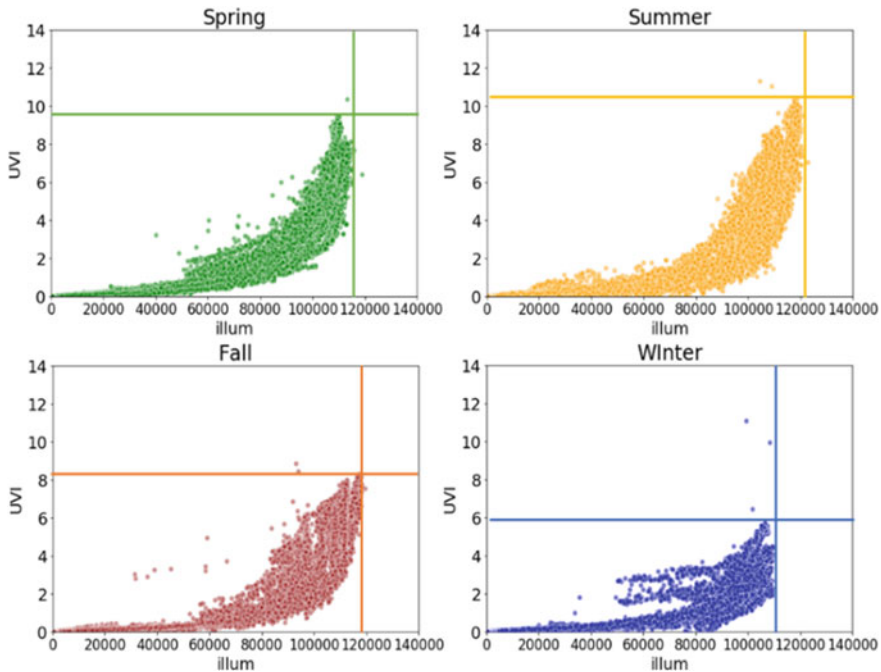


Fig. 1 Illuminance—UVI scatter plots by season

Figure 1 presents UVI distribution according to the illuminance values by extracting the data on the clear days in each season. The distribution of UVI by season shows a different trend. The maximum illuminances during spring, summer, and autumn are similar with about 120,000. While, the maximum UVI during spring and autumn is 8–9, and it is higher than 10 during summer. Especially, the UVI under the same illuminance is widely distributed. In addition, during winter, the maximum illuminance is about 110,000 and maximum UVI was as low as lower than 6. The analysis results indicated that the correlation between illuminance and UVI by season is different, expecting that it would be also advantageous to reflect seasonal characteristics of natural light in the calculation of illuminance-based UV information in the future.

3 Deep Neural Network Model for Calculating Illuminance-Based UV Information

The deep neural network model for the calculation of illuminance-based UV light information uses the illumination value as the main input variable and the solar zenith angle that is found to be highly correlated with UVI. In addition, monthly information was added as an input variable to reflect the seasonal characteristics of UVI. Figure 2 shows the construction of the deep neural network model proposed in this paper.

In the deep neural network model in Fig. 2, the input variable was set a total of 15 including illuminance, solar zenith angle, and seasonal characteristics data. While, initialization of weight is set to arbitrary value of standard normal distribution. The scale of illuminance was adjusted to 10,000:1 for smooth learning. In addition, the seasonal characteristics were categorized into monthly information in order to be

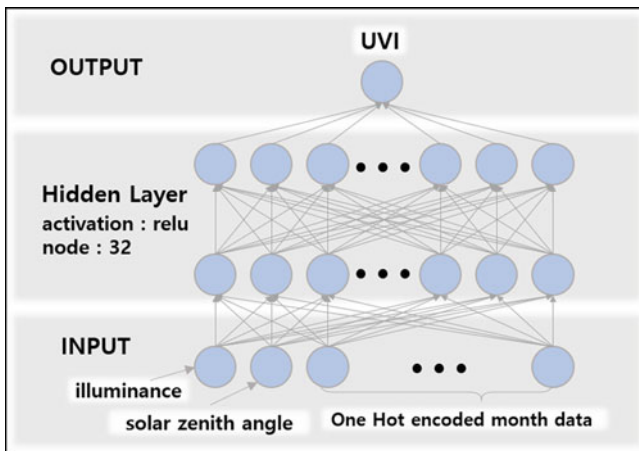


Fig. 2 Schematic diagram of deep neural network model

reflected as a more detailed seasonal characteristics before applying them as the input variable. At this time, each weight of month is assigned by using one-hot encoded 13-bit data to input monthly information corresponding to categorical data instead of general integer variables. After that, the learning data was constructed through two hidden layers. Each node consisted of 32 nodes and the active function was constructed with ReLU. Furthermore, the loss function of the model is set to MAE (average absolute error). ADAM was adopted as an optimization algorithm and MSE (mean square error) was adopted to calculate one output UVI through learning the deep neural network model.

4 Validation and Analysis

In order to validate the proposed deep neural network model, first the performance according to application of the seasonal characteristics data was compared. The results are tabulated in Table 2.

Table 2 presents the results when the natural data for two years were learned. When the monthly data of seasonal characteristics were applied along with the illuminance and solar zenith, the test performance was 0.29 as per MAE and 0.25 as per MSE, showing a relatively low error. In addition, an experiment was performed to check calculation performance of the illuminance based UVI of the proposed deep neural network model. Arbitrary data were extracted per season for clear days and cloudy days one day each that were not included in the learning. These data were used in the two models constructed as above as input variables to validate performance difference. The results are shown in Fig. 3.

In Fig. 3, errors are obtained as mean value at each hour from the UVI output values in the deep neural network model and compared the mean of each error (MAE) by season. In the model without reflecting the seasonal characteristics, the error was relatively low in the winters. However, the error rates were high for other seasons. Whereas, when the seasonal characteristics were considered in the learning, constantly low errors were found in the spring, summer, and autumn except winters. The comparison results for entire MAE showed that both models yield errors of 0.25 and 0.17, respectively. The results confirmed that the proposed deep neural network model with the seasonal characteristics can calculate more accurate UVI.

Table 2 Error according to application of monthly data

Input variables	MAE	MSE
Illuminance, solar zenith angle	0.37	0.39
Illuminance, solar zenith angle and seasonal characteristics data	0.29	0.25

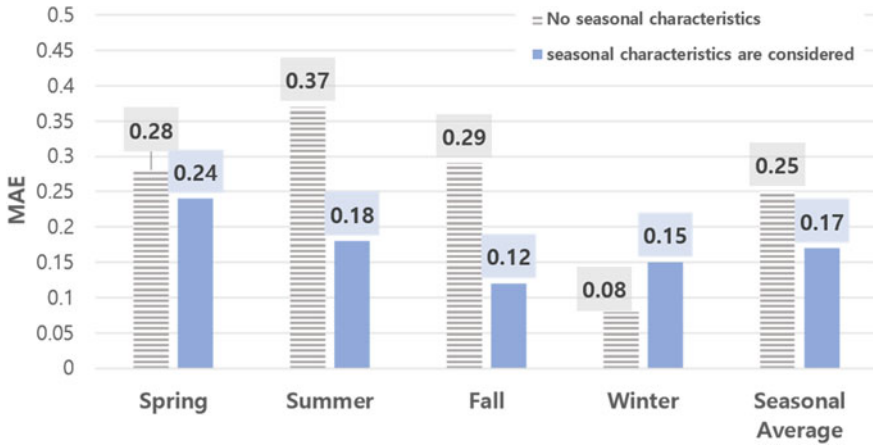


Fig. 3 Comparison of error in the deep neural network model

5 Conclusion

In this paper, a deep neural network model in which seasonal characteristics of natural light are reflected is proposed to calculate an illuminance-based ultraviolet information. The correlation between the light characteristics was analyzed by using measured natural light data. The analysis results showed that UVI, illuminance, and solar zenith were highly correlated. In addition, the illuminance and UVI showed different correlation according season, indicating that seasonal characteristics of natural light need to be reflected in the deep neural network model. The learning dataset was then extracted from the measured natural light data. The deep neural network model to calculate the illuminance based UVI was then established by applying 2 hidden layers, 32 nodes in each stage, and activation function ReLU. Into the input data, one-hot encoded monthly data were also added in order to reflect seasonal characteristics of natural light along with the illuminance and solar zenith, and then output was set as UVI values. The performance of the proposed deep neural network model before and after applying the seasonal characteristics were compared and analyzed. The proposed model with input data of illuminance, solar zenith, and seasonal characteristics showed a lower error (MAE) by 0.08 than that of model with only illuminance and solar zenith data, confirming the proposed model has a high degree of accuracy in calculating UVI.

In the future, additional learning and experiments would be needed to improve performance of the proposed model in order to calculate more accurate UVI for all the seasons. In addition, a research would be performed to establish a system based on the proposed model which not only provides UVI but also UV related information such as erythema weighted ultraviolet light (EUUV) and vitamin D to the users by linking illuminance and locational information from smart phones.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A2B2005601).

The research was supported by the International Science and Business Belt Program through the Ministry of Science and ICT (2015-DD-RD-0068-05).

References

1. World Health Organization, International Commission on Non-Ionizing Radiation Protection (2002) Global solar UV index: A practical guide. World Health Organization
2. Kim HM, Oh ST, Lim JH (2018) Development of local area alert system against particulate matters and ultraviolet rays based on open IoT platform with P2P. Peer-To-Peer Netw Appl 11(6):1240–1251
3. KM Kim 2017 Design of lighting system to support vitamin D synthesis using natural light measurement data. Master's dissertation, KongJu National University Republic of Korea Cheonan

Model for Classifying Color Temperature Anomalies of Natural Light in Real Time Using Deep Learning



Geon-Woo Jeon, Seung-Taek Oh, Heon-Tag Kong, and Jae-Hyun Lim

Abstract Modern people spend longer time in artificial lighting environments, it causes problems such as biorhythm imbalance. In order to resolve this problem, studies are being conducted to reproduce the characteristics of natural light through indoor lighting. However, researches for detect anomalies of color temperature caused by cloud and weather changes during real-time measurement of the color temperature are lacking. This paper proposes classification algorithm of real-time color temperature anomalies of natural light based on deep learning. First, anomaly was defined by analyzing daily color temperature pattern on the clear day and cloudy day with the natural light data measured for two years. The factors related with the occurrence of anomalies were explored and a data set to be used in the deep learning model was extracted. The deep learning model that uses LSTM layer which is beneficial to reflect time series characteristics of the extracted data was designed.

Keywords Anomaly classification · Natural light · Deep learning

G.-W. Jeon (✉) · H.-T. Kong · J.-H. Lim

Department of Computer Science and Engineering, Kongju National University, Cheonan, Republic of Korea

e-mail: momentum96@smail.kongju.ac.kr

H.-T. Kong

e-mail: htkong@kongju.ac.kr

J.-H. Lim

e-mail: defacto@kongju.ac.kr

S.-T. Oh

Smart Natural Space Research Center, Kongju National University, Cheonan, Republic of Korea

e-mail: ost73@kongju.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_28

1 Introduction

Periodic characteristics of natural light from sunrise to sunset help in maintaining human physiological functions such as circadian rhythm and sleep cycle [1]. However, indoor artificial lighting environment that provides fixed illuminance or color temperature can disturb circadian rhythm and sleep cycle [2]. In order to solve this problem, studies are continued to reproduce the characteristics of natural light through indoor lighting [3]. However, previous studies have only reproduced some characteristics of light or periodic characteristics of the collected natural light, thus could not reproduce the characteristics of natural light that is changing every moment with day, month, and year-cycles. In addition, in order to reproduce natural light in real-time, the natural light which is periodically changing at every moment need to be measured. The researches were conducted to measure the color temperature in real-time, but anomalies of the measured color temperature that can occur when the color temperature is abruptly changed due to cloud or weather could not be investigated. Therefore, this paper proposes a classification model for the real-time color temperature anomalies based on the deep learning to reproduce natural light in real-time. For that, daily changing patterns of the color temperature and light characteristics data on the clear days and cloudy days each based on the natural light database measured over two years were analyzed and anomaly value of color temperature was defined. In addition, the factors related with the occurrence of the color temperature anomalies were derived and data set to construct anomalies classification algorithm was prepared. And then a deep learning model wherein LSTM layer was used was designed so that time series characteristics of the input data can be reflected in the proposed model. The leaning with the data set composed of 80% of whole natural light data was performed and anomalies classification algorithm was evaluated by using data for one day in which color temperature anomaly is existed.

2 Characteristics of Natural Light and Anomalies of Color Temperature

2.1 Characteristics Analysis of Measured Natural Light

In this study, the characteristics data of sunlight were collected by tracking the sun at one-minute interval from the roof of a 10-story building located at latitude 36.85 and longitude 127.15 from April 2017 to May 2019. The database for the natural light was established with the collected data. The equipment for the measurement was a spectroradiometer (CAS140-CT, Instrument Systems) that measures the spectrophotometric illuminance of light to calculate and provide illuminance, color temperature, wavelength ratio, and UV intensity. First, in order to investigate the changing characteristics of light including color temperature for clear day and cloudy day, the characteristics of the light for the days with average cloudiness of 0.0 and 4.5



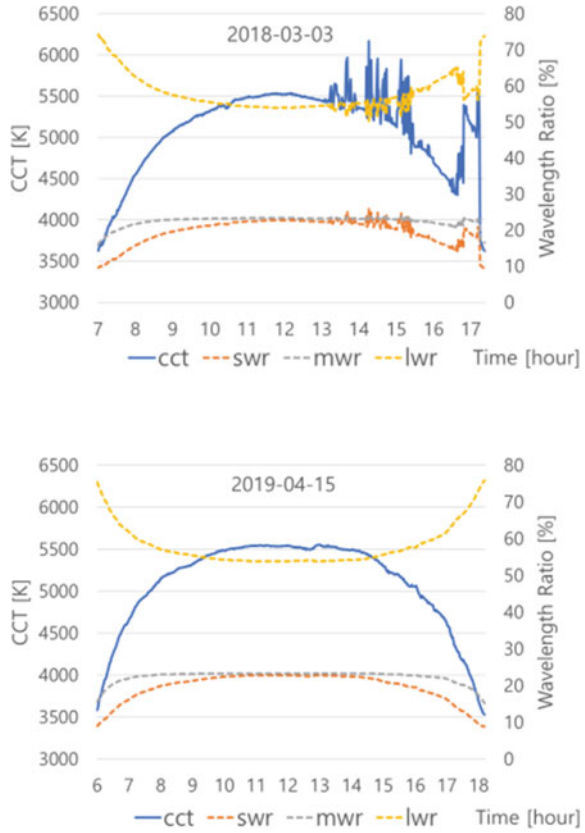
Fig. 1 Illuminance measurement data for the clear day and cloudy day

among the natural light database were analyzed. Figure 1 shows the daily illuminance distribution for the selected clear day (Apr. 15, 2019) and cloudy day (Mar. 3, 2018).

In Fig. 1, the illuminance of clear days was relatively higher than those of cloudy days, and the illuminance distribution of clear days exhibits a parabolic form with the illuminance constantly increasing and decreasing. However, the illuminances of the cloudy days were lower than those in the clear days by on an average 20,000 lux and the illuminance distribution was also uneven. It might be due to sunlight that is transmitted was reduced due to clouds or was abruptly changed. The changes of the color temperatures and wavelength ratios during the selected clear days and cloudy days are presented in Fig. 2.

In Fig. 2, the color temperatures during the clear days and cloudy days show similar distributions as in Fig. 1. Furthermore, in the graph for the clear day in Fig. 2, the color temperature and ratio of the short wavelength (380–480 nm), medium wavelength (480–560 nm), and long wavelength (560–780 nm) are evenly distributed. Whereas, in case of the cloudy day, the wavelength ratio was irregular in the zone where the color temperature was measured irregular. Yet, the change was relatively less than that of CCT, indicating that the wavelength ratio was not affected by the clouds or weather as much as the illuminance or color temperature. However, cloudy day in Fig. 2, the short wavelength was instantly increased due to clouds’ blocking the sun during 14:00–16:00 and 17:00–17:30, and it was changed in every minute. It might be because, the light of red series long wavelength was blocked by the clouds in a short duration and the light of blue short wavelength reached more to the measurement site. The analysis results for the characteristics of natural light on clear days and cloudy days confirmed that the characteristics of illuminance and short wavelength ratio of natural light may be similar to color temperatures and distribution characteristics.

Fig. 2 Color temperature and wavelength ratio: clear days and cloudy days



2.2 Characteristics Analysis of Measured Natural Light

In order to define the color temperature anomalies necessary for evaluating the real-time color temperature anomalies, natural light data were analyzed. For that, entire natural light data and the distribution of color temperature on the clear days having constant color temperature distribution were analyzed and summarized. For the data of clear days (cloudiness 0.0), one day per month was selected and the color temperature values measured at one-minute interval from sunrise to sunset for 12 days were used. While, the color temperature for all the days in the natural light database which are not categorized by clear and cloudiness of the day were used as the total data. In addition, after calculating the changed amounts (x_2-x_1 , Δx) of two neighboring color temperature values measured at one-minute interval, the distribution status of the resulting value (Δx) was analyzed and the results are tabulated in Table 1.

Analysis results showed that more than 99% color temperature change data were distributed in the zone of higher than $-50-50$. In addition, the distribution of the color temperature change for whole data confirmed that color temperature change

Table 1 Changed amount distribution of the measured color temperatures

Changed amounts (Δx) range	Changed amount distribution	
	Clear days (%)	All days (%)
$-30 \leq \Delta x \leq 30$	97.26	81.3
$-40 \leq \Delta x \leq 40$	98.14	85.1
$-50 \leq \Delta x \leq 50$	99.40	87.3
$-100 \leq \Delta x \leq 100$	99.58	92.3
$-200 \leq \Delta x \leq 200$	99.7	95.6

data are presented more than 90% in the zone higher than $-100-100$ and 87.3% in the zone lower than $-50-50$. Furthermore, the color temperature change zone in $-50-50$ accounted more than 87% of the entire color change amount and accounts almost all the data (more than 99%). Therefore, it is the zone from where the color temperature of the clear days can be distinguished easily. Based on the above results, the zone of $-50 \leq \Delta x \leq 50$ was set as a reference where the measured color temperature anomalies can be tracked. When the color temperature change became an outlier from the zone, that color temperature was defined as an anomaly of the measured color temperature.

3 Anomalies Classification Model for the Color Temperature of the Real-Time Natural Light

The deep learning-based color temperature anomalies classification model was constructed as three stages: input data construction and preprocessing, deep learning model design, and model evaluation and analysis. First, pretreatment for the natural light data was proceeded to process them into data that can be input into the deep learning model. The input data for the color temperature anomalies classification model were prepared with the illuminance and short wavelength ratio that were confirmed as related with the color temperature anomalies through characteristics analysis of the measured natural light. The incidental information such as relative duration and short wavelength ratio at one-minute interval were also included in the input data of the color temperature anomalies classification model. The relative duration was scaled as 0 when the color temperature was at minimum after sunrise, and as 1 when the color temperature was minimum before sunset. The data standardization which converts average value to 0 and standard deviation to 1 was applied in the short wavelength ratio and illuminance data. In addition, the data set for leaning the model were extracted randomly at 80% of the whole natural light data. The deep learning model for real-time color temperature anomalies classification consists of a LSTM layer with 128 nodes, a hidden layer with 256 nodes, and an output layer with one neuron. Furthermore, for input data, the previous 10 numbers of data collected at one-minute interval which include the moment of anomalies classification were used.

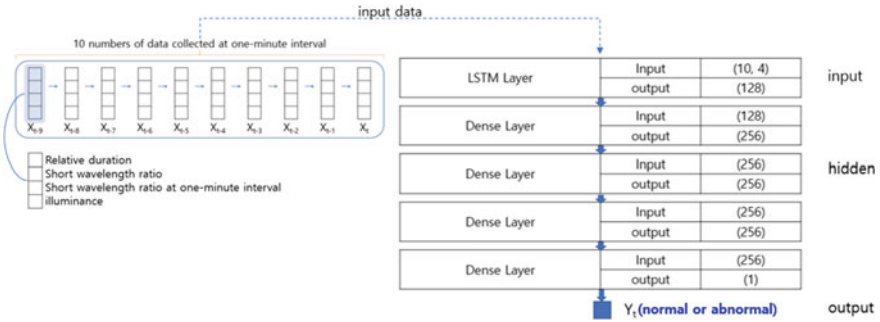


Fig. 3 Schematics of model and input data

Each data was constructed with the relative duration, short wavelength ratio, short wavelength changes amount, and illuminance. Figure 3 shows the schematic diagram of anomalies classification model and input data and t indicates the measurement moment.

When building the deep learning model shown in Fig. 3, three dense layers adjacent to the LSTM layer used ReLU, which is advantageous for error back propagation when training the model. In the dense layer for final output, a sigmoid activation function was applied to make the final output have probability between 0 and 1. In addition, the model’s loss function uses binary cross entropy provided by Keras library to achieve binary classification purposes. The optimizer used for optimization was Adam Optimizer. The batch-size which is the sample size for weight update was set to 64 and the maximum number of training round was set to 100.

In order to evaluate the performance of the proposed model, the color temperature prediction results were monitored by using the color temperature data of one day (March 3, 2018) that was not used in the training. Figure 4 shows the distribution of color temperature for the day used in the evaluation, and the red-colored area shows the result of detecting anomalies of color temperature which may be caused by the effects of cloud and weather. When the non-anomalies are classified as normal values, it is confirmed that the anomalies and normal values of color temperature can

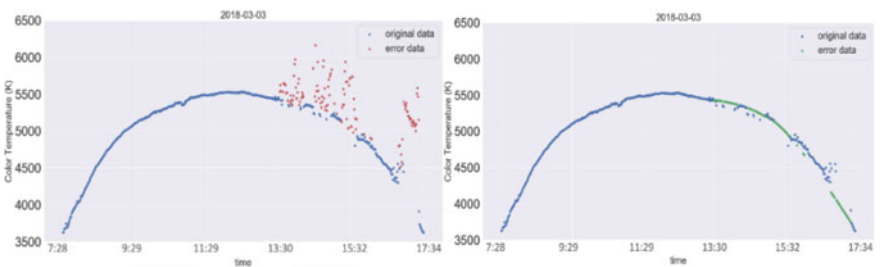


Fig. 4 Execution of the real-time color temperature anomalies classification model and Replacement of detected anomaly color temperature

be classified with a probability of 98.87%. In addition, in order to reproduce the periodic characteristics of natural light, actual measurement days of natural light having a pattern like the color temperature distribution were selected and results of replacing some of the color temperature of that day with the anomalies at Fig. 4.

Figure 4 shows replacement results of the anomalies with the color temperature value of natural light on March 8, 2019. This enables the real-time cyclic characteristics of the natural light to be reproduced when the corresponding part is replaced with the normal color temperature after detecting the color temperature anomalies through the proposed model.

4 Conclusion

In this paper, a deep learning based real-time color temperature anomalies classification model is proposed. The natural light data measured over two years were analyzed to confirm that illuminance and short wavelength ratio are factors related to color temperature anomalies. In addition, the color temperature anomaly value is defined as the value when the color temperature variation was more than -50 to 50 K than that measured moment before. In addition, a color temperature anomalies classification model was designed to classify status of color temperature anomalies by entering the illuminance, short wavelength ratio, amount of short wavelength changes at one-minute interval, and relative time. The proposed model was constructed with the one layer of LSTM, 4 layers of dense layer. The leaning was performed by using 80% of the measured natural light data. The model was evaluated and validated by using the light characteristics data on a specific day (March 3, 2018) when the cloudiness was 4.5. The validation results showed that the anomalies and normal values of the color temperature can be classified with the probability at 98.87%. When the anomaly values of the color temperature of natural light is judged through the proposed model and replaced it with the appropriate color temperature, it is possible to reproduce the periodicity of natural light in real-time. In the future, research is needed to develop and implement measurement devices that can easily measure the illuminance and short wavelength ratio to be used as inputs for the proposed model. In addition, research is planned to develop a color temperature matching algorithm that replaces the color temperature anomalies with the normal values to reproduce the periodic characteristics of natural light.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A2B2005601), This work was supported by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1A6A1A03032988).

References

1. Roenneberg T, Merrow M (2007) Entrainment of the human circadian clock. Cold Spring Harbor symposia on quantitative biology, vol 72. Cold Spring Harbor Laboratory Press, pp 293–299
2. Stevens, RG et al (2014) Breast cancer and circadian disruption from electric lighting in the modern world. *CA: A Cancer J Clin* 64(3):207–218
3. Kim K et al (2019) Development of a natural light reproduction system for maintaining the circadian rhythm. *Indoor Built Environ* 0(0):1–13

Low-Resolution LiDAR Upsampling Using Weighted Median Filter



Hyun-bin Lim, Eung-su Kim, Pathum Rathnayaka, and Soon-Yong Park

Abstract This paper presents a 3D LiDAR data upsampling method to obtain dense 3D depth data from a low-resolution LiDAR, a vision camera, and the Weighted Median Filter (WMF) algorithm. Recently, LiDAR is widely used in the field of Computer Vision since it can obtain accurate 3D data. However, data acquisition from the LiDAR is expensive due to the high cost of the LiDAR. We address how to obtain large amounts of 3D data from a low-channel LiDAR. In this paper, we acquire LiDAR data and color images from a calibrated multi-sensor platform. We first begin the upsampling steps from linear interpolation of a depth image. And then, we use an WMF algorithm to complement the first interpolation image. We use the upsampling algorithm to create dense 3D depth map from the existing sparse LiDAR data. And we generated a high-density 3D map using the ICP matching of multiple depth data acquired by a moving robot system.

Keywords LiDAR · Upsampling · Weighted median filter · Guided image filter · ICP

H. Lim · E. Kim · P. Rathnayaka

School of Computer Science & Engineering, Kyungpook National University, Deagu, South Korea

e-mail: limgusqls@gmail.com

E. Kim

e-mail: jsm80607@gmail.com

P. Rathnayaka

e-mail: bandarapathum@yahoo.com

S.-Y. Park (✉)

School of Electronics Engineering, Kyungpook National University, Deagu, South Korea

e-mail: sypark@knu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_29

1 Introduction

Recently, LiDAR can be used to a variety of Computer Vision areas, such as object detection, recognition, and 3D map generation because LiDAR can obtain accurate 3D data. Therefore, LiDAR data is getting important. In addition, as LiDAR data becomes more important, the amount of LiDAR data becomes more important. Because, the greater the amount of LiDAR data, the better the research results. For this reason, high-resolution LiDAR is often used to obtain large amounts of LiDAR data. However, the larger the number of LiDAR channels, the more expensive the LiDAR device is, which causes cost problems. Solve this cost problem, researches [1, 2] for upsampling low channel LiDAR data are being conducted.

Two methods of LiDAR upsampling are introduced in [1, 2]. Both upsampling methods use LiDAR and color images. And, both methods upsampling using the depth map. Wirges et al. [1] present an improved model for MRF-based depth upsampling, guided by image-as well as 3D surface normal features. And [2] use Synchronized color image and Anisotropic Diffusion Tensor to upsampling. The difference between the two studies [1, 2] and our study is the number of LiDAR channels used in the study. Two research used 64-channel LiDAR for upsampling, but we used 16-channel LiDAR for upsampling.

Our research presents a method for upsampling low-resolution LiDAR data using low-resolution LiDAR data and vision camera image data. First, we calibrate LiDAR and the vision camera. Because calibration can be used to create a depth map with the same resolution as the vision camera image. We upsampling LiDAR data using a depth map containing the distance data of the LiDAR. Before applying the weighted median filter, linear interpolation is performed first. Then, to compensate for the interpolated depth map, the final upsampling is performed by applying a weighted median filter. Through this process, the previous sparse LiDAR data can be transformed into dense 3D data. Details of each method will be explained in detail in Session 2.

2 LiDAR Depth Upsampling

2.1 LiDAR and Camera Calibration

Our research use both the color data of the camera and the 3D point of LiDAR. First, we need a camera and LiDAR calibration. Because camera calibration can be used to determine the relative transformations of color image data and LiDAR data. In addition, calibration can make a depth map and colored LiDAR data. Various LiDAR and camera calibrations exist. However, most methods require the use of a special calibration board. In order to easily calibrate, we used the [3] method. Because [3] is a method of calibrating only using chessboard.

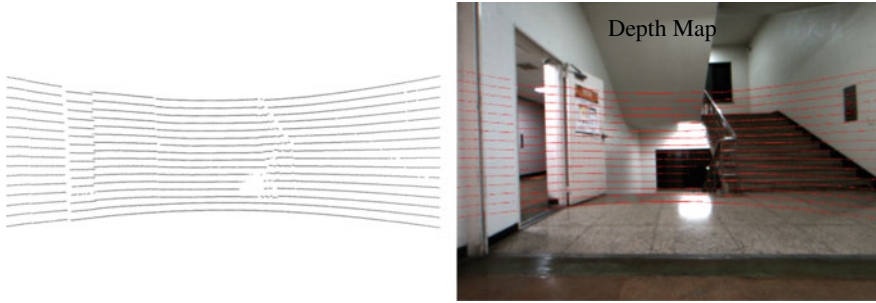


Fig. 1 Depth map and color image, The black points in the depth map (top image) each contain a LiDAR distance value. The red points in the Color Image indicates the point that corresponds to the black points in the depth map

2.2 Depth Map

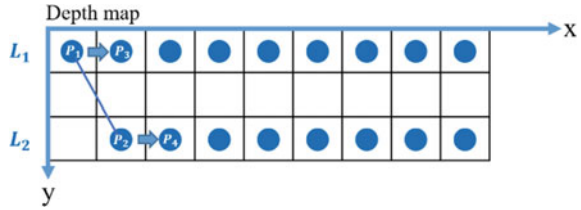
We use the depth map for upsampling LiDAR data. The depth map is an image or image channel that contains information relating to the distance of the surfaces of scene objects from a viewpoint. We have done camera and radar calibration in an advanced step. Calibration can get the relative transformation of LiDAR and color images. So, we can use relative transformation to get a depth map that is the same size as the color image. The first image in Fig. 1 is the depth map and the second image shows where the LiDAR data is projected onto the color image.

2.3 Four Point Selection

We interpolate using the depth map. In order to proceed with interpolation, we must select four points of LiDAR points. Before selecting four points, the channels of LiDAR points should be classified according to the altitude. Each channel can be distinguished because it has the same altitude.

Each channel is called $L_i (i = 1, 2, 3, \dots, 16)$. And four points are $P_k (k = 1, 2, 3, 4)$. First, select P_1 from L_i . Second, P_2 having an azimuth equal to P_1 is selected among the L_{i+1} channels that are next to the L_i channel. Third, select the next azimuth point adjacent to P_1 as P_3 . Fourth, the next azimuth adjacent to P_2 among the L_{i+1} channel data is selected as P_4 . The way to select four points is easy to understand with Fig. 2.

Fig. 2 Select four points. P_1 and P_2 are points with the same azimuth, and P_3 and P_4 are points with azimuths adjacent to P_1 and P_2



2.4 Depth Interpolation

Perform a linear interpolation on the depth map using the four selected points. D_k is the pixel distance from four points to the point to be newly interpolated. W_k is the weight of each point. P' is a newly interpolated point. P_k is the z data of the LiDAR point. The meaning of each parameter can be understood from Fig. 3. And the depth interpolation formula is as below.

$$W_k = e^{-\lambda D_k} \tag{1}$$

$$P' = \frac{\sum_{k=1}^4 W_k P_k}{\sum_{k=1}^4 W_k} \tag{2}$$

The weight of the distance from four points to the newly interpolated point is calculated by Eq. (1). And the new point P' is obtained from Eq. (2) by applying weights to four points. Interpolation results are shown in Fig. 4. The left image in Fig. 4 is the depth map before interpolation and the right image is the interpolated depth map.

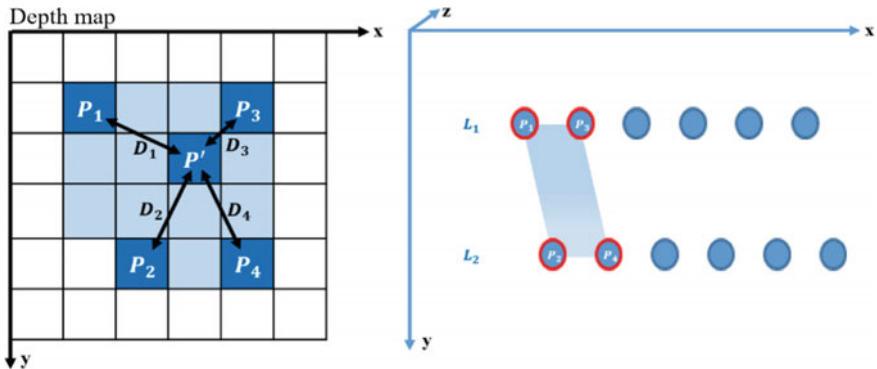


Fig. 3 Four points interpolation. D_k is the pixel distance. P_k . In the second image, the blue area shows the interpolated area

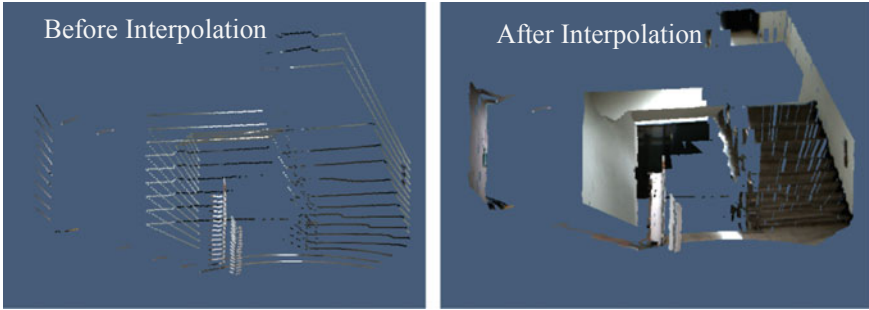


Fig. 4 3D points before and after interpolation Compared with the left, the right has significantly increased 3D data

2.5 *Weighted Median Filter*

Weighted Median Filter (WMF) [4, 5] is a filtering method that utilizes Guided Image Filter [6]. Guided Image Filter is Edge Preserving Filter proposed by He et al. [6]. Guided Image Filter is both effective and efficient in a great variety of computer vision including edge-preserving smoothing, detail enhancement, HDR compression, image matting, feathering, dehazing, joint upsampling. Guided Image filter performs Edge preserving filtering using color information of Guide Image. We use the edge-preserving smoothing effect of the Guided Image Filter. Using a color image acquired with a vision camera as a guide image, and the Guided Image Filter is performed on a depth map that is the same size as the color image. By using edge-preserving smoothing, the parts that are not interpolated can be compensated by the Guided Image Filter.

WMF is an improved algorithm of Guided Image Filter. Since we show better results than the Guided Image Filter, we use the WMF to improve the interpolation depth map. WMF is an algorithm that divides the filtering image into layers by distance, applies a Guided Image Filter to each layer, and then applies a median filter. We used the color image acquired through the vision camera as a guided image, and the filtering image used the interpolated depth map. In the experiment, we use WMF that exists in OpenCV, an open-source library. OpenCV WMF divides the filtering range into 256 layers and applies a Guided Image Filter to each layer. So we did the WMF by dividing the depth map by 256 cm intervals for apply 1 cm intervals Guided Image Filter. The process can be understood through Fig. 5. And the first image in Fig. 6 is the guide image, the second image is the interpolated depth map, and the third image is the filtered depth map. If you compare the second image and the third image in Fig. 6, you can see that more data is generated.

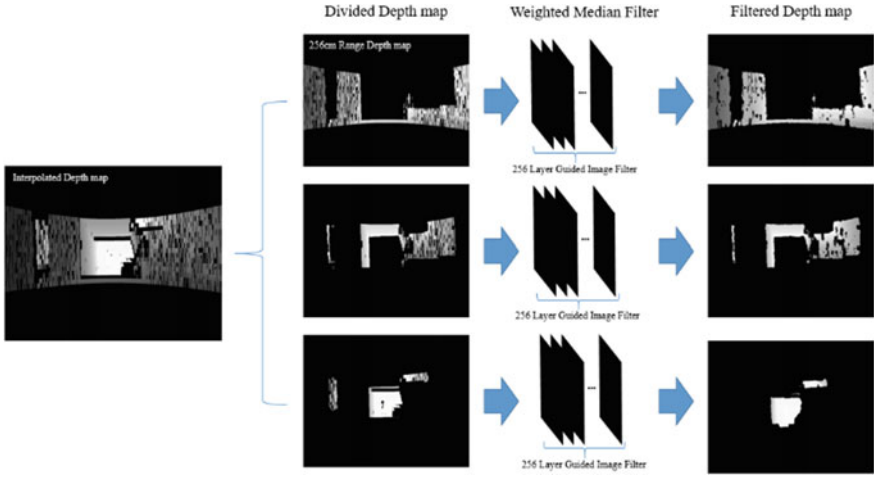


Fig. 5 Weighted Median Filter process. The algorithm proceeds from left to right. First, we divide the depth map by 256 cm. Then use Weighted Median Filter for each divided depth map. Finally we can get the filtered depth map

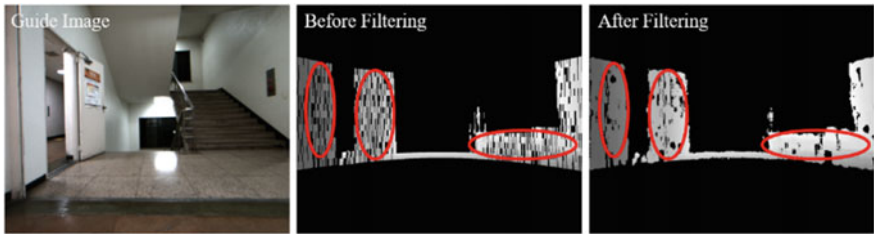


Fig. 6 Weighted Median Filtered Depth map. First image is the guide image, the second image is the interpolated depth map, and the third image is the filtered depth map

3 Experiments

We make a multi-sensor for the fusion of LiDAR and camera to collect data. LiDAR is mounted on the top and two cameras are mounted on the bottom. LiDAR used Velodyne VLP16. And camera used FLIR Flea3 vision camera. And lens used 3.5 mm lens. In the experiment, the data was collected by mounting the multi-sensor on the robot. The robot model is Pioneer P3-DX.

We obtain LiDAR data and color images 20 degrees left and right through multi-sensor. We created a 3D map using the upsampling results to check the upsampling result. In our research, we experimented with the Iterative Closest Point (ICP) algorithm provided by Point Cloud Library (PCL) as a simple way to create 3D maps. ICP algorithm is a method of registering current data in an existing data set. It uses the closest point of each data to find an association and move and rotate the current



Fig. 7 Depth upsampling result (left) and 3D map building result (right)

data accordingly. An experimental result after ICP-based mapping is shown in Fig. 7. The left two images are before and after upsampling results. At the right of the figure is the front view of the 3D map made with ICP. In our experiments, we were able to acquire a much larger amount of data than existing LiDAR data. And we can get the high-density 3D map using the ICP algorithm.

4 Conclusion

We presented a LiDAR depth data upsampling method from a low-channel LiDAR sensor and a vision camera. A multi-sensor setup is used in the experiment and it is calibrated. First, we calibrate multi-sensors to get depth maps. Second, we interpolate the depth map. Afterward, the weighted median filter is used to compensate for the depth map data that could not be filled by linear interpolation. As a result, we can see from Fig. 7 that the amount of existing low-channel LiDAR data has increased greatly through upsampling. And we created a dense 3D map using a simple ICP algorithm.

Acknowledgements This research was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT: Ministry of Science and ICT) (No. 2018M2A8A5083266).

Reference

1. Wirges S, Roxin B, Rehder E, Kühner T, Lauer M (2017) Guided depth upsampling for precise mapping of urban environments. *Intell Veh Symp (IV)*, pp 1140–1145
2. He Y, Chen L, Li M (2015) Sparse depth map upsampling with RGB image and anisotropic diffusion tensor. *Intell Veh Symp (IV)* 205–210
3. Kim ES, Park SY (2019) Extrinsic calibration of a camera-LIDAR multi sensor system using a planar chessboard. In: *Eleventh international conference on ubiquitous and future networks*, pp 89–91
4. Ma Z, He K, Wei Y, Sun J, Wu E (2013) Constant time weighted median filtering for stereo matching and beyond. In: *International conference on computer vision*, pp 49–56
5. Zhang Q, Xu L, Jia J (2014) 100+ times faster weighted median filter (WMF). In: *Conference on computer vision and pattern recognition*, pp 2830–2837
6. He K, Sun J, Tang X (2012) Guided image filtering. *IEEE Trans Pattern Anal Mach Intell* 6:1397–1409

Design of Tablet-Based Live Mobile Learning System Supporting Improved Annotation



Jang Ho Lee

Abstract In the past, we had developed a mobile learning system that delivers a lecture to students in the distance through tablets. The students were able to watch the lecture with presentation slides with annotation as well as to ask questions in chat with their tablets in real time. An instructor, however, sometimes had difficulty explaining a concept just by making annotations in the panel for presentation slides since there is usually not enough space to draw figures necessary for explanation by hand in real time in that panel. Therefore, we present a design of a newly improved system with a separate whiteboard panel in which an instructor can draw figures by hand in real time so that the students can better understand the concept being explained by the instructor.

Keywords Tablet · Live mobile learning · Annotation · Whiteboard

1 Introduction

Recent popularity of mobile devices has made the distance learning system based on a smartphone or a tablet draw a tremendous attention from researchers [1].

One of the pioneers of tablet-based learning system is Classroom Presenter [2]. With this system, students were able to share slide as well as annotation with a teacher using their tablets. However, the system was only used with a teacher and students physically present in the same classroom since it did not support video nor chat for real-time interaction for people who are geographically apart.

And one of the early live mobile learning systems is MLVLS [3]. The system allowed students to watch not only lecture video but also presentation slide with annotation on a smartphone in real time. However, the size of the smartphone display was small for the students to see the slide comfortably and there is not interaction capability between a teacher and students.

J. H. Lee (✉)

Department of Computer Engineering, Hongik University, 72-1 Sangsu, Mapo, Seoul 121-791, Korea

e-mail: janghol@hongik.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_30

We had also developed a live mobile learning system in the past [4]. It allowed students to watch a lecture video and presentation slide with annotation being made by an instructor in real time. Students can also ask questions with chat on a tablet in real time. Although the annotation can be made on a slide panel in this system, there was usually not enough empty space in the slide page for drawing a figure in real time to help students better understand the concept in the slide page. In the real conventional classroom lecture with an instructor and students in the same room without any mobile learning system, this type of problem would not have occurred because, in addition to the projector screen showing presentation slide, there is usually a blackboard or whiteboard for drawing figures.

Therefore, we propose the design of a tablet-based live learning system with whiteboard capability for improved annotation in order to help students better understand the lecture. In this newly-proposed system, if the instructor needs to explain the concept with some figure that is not in the presentation slide but the empty space in the presentation slide is too small for the figure to fit in, he can draw this figure on the whiteboard that are shown on the whiteboard panel in the UI of the iPad client. We hope this will help raise the degree of students' understanding of the lecture close to the level of real classroom environment.

2 Tablet-Based Live Learning System with Whiteboard to Improve Annotation

The concept of the presented tablet-based live learning system with whiteboard is shown in Fig. 1. An iPad tablet is used for the client both for instructor and for students. An instructor gives a lecture in front of his iPad tablet, the video and audio of the lecture are fed to the tablets of the students who are geographically apart via LTE/WiFi in real time. Students can also watch the slide with annotation that are

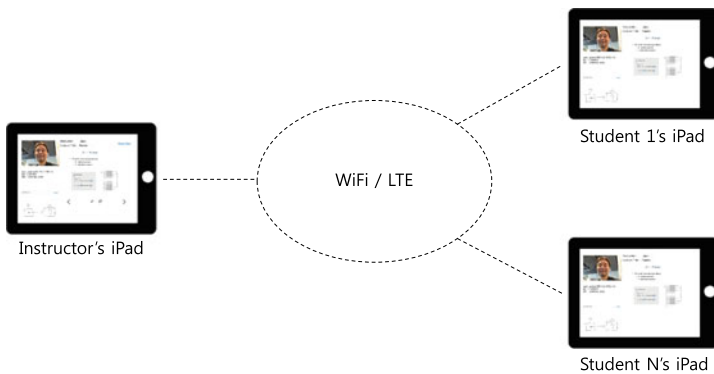


Fig. 1 Table-based live learning system with whiteboard to improve annotation

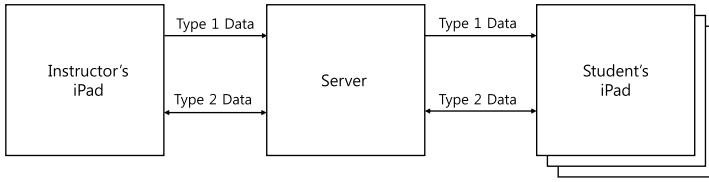


Fig. 2 Communication architecture of the tablet-based live learning system with whiteboard to improve annotation (One instructor’s iPad and multiple student’s iPads)

being made by the instructor in real time. They can ask questions by typing messages on the chat panel, which can be answered by the instructor via video, audio, slide with annotation, and text chat.

Since the empty space for annotation on a slide panel is limited and too small for a figure that the instructor needs to draw instantly to help students to understand the lecture, we newly added a whiteboard panel in our tablet client.

Figure 2 illustrates the communication architecture of the proposed tablet-based live learning system with whiteboard to improve annotation. The basic components are tablet clients and a server. There are two type of tablet client. One is the instructor’s iPad client and the other is the student’s iPad client. There is only one instructor iPad client while student’s iPad client can be more than one.

Type 1 data is the data that are only sent from instructor’s tablet and then multicast to all the iPad clients. Type 1 data includes video, audio, slide, annotation being drawn on the slide, and the drawing data on the whiteboard. On the other hand, Type 2 data is the data that can be sent from any iPad and then multicast to all the iPads through server. Type 2 data includes text message on a chat, and session update message (e.g., joined a session, left a session).

About the architecture of streaming of video and audio of the instructor’s iPad to the students’ iPads through server, HTTP Live Streaming (HLS) architecture [5] was chosen. In this HLS architecture, the video and audio input is encoded as HEVC video and AC-3 audio and output to a MPEG-4 format which is broken into a short media files and placed on a web server with an index file containing a list of media files so that the client reads the index and the listed media files.

The prototype user interface of the instructor’s iPad client is illustrated in Fig. 3.

The instructor starts a lecture by clicking “Open files” button and choose the appropriate lecture file and enter the lecture title. While giving a lecture, the instructor can move to the next slide or to the previous slide using slide control buttons at the bottom of the slide panel. He can also make annotations on a slide. When an instructor needs to draw some figure in real time to help students better understand the slide, he can do it by using the whiteboard panel which is lower left side of the UI in Fig. 3.

The prototype user interface of the student’s iPad client is shown in Fig. 4.

The students can watch the instructor’s gesture and listen to his voice. They also see the presentation slide with annotation made by the instructor in real time. The

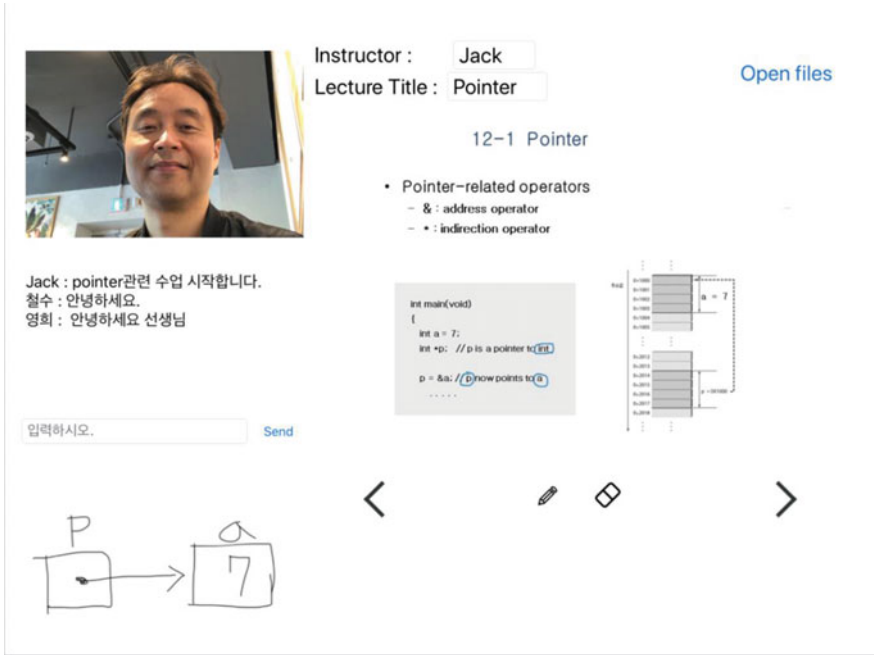


Fig. 3 User interface of the instructor’s iPad

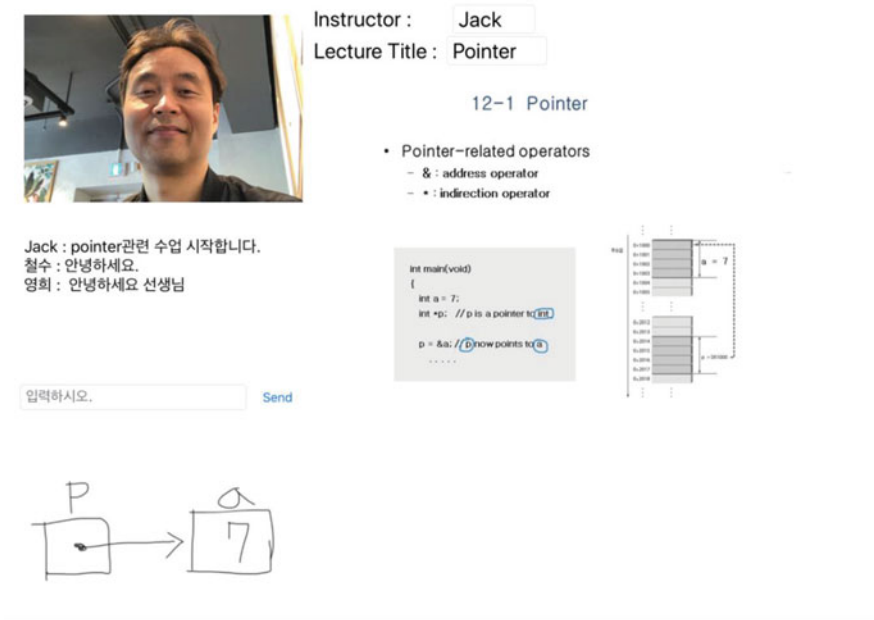


Fig. 4 User Interface of the student’s iPad

students can also ask questions by typing texts on a chat, which are shown at the bottom of the video panel of UI. When a question is asked, the instructor can answer by using voice, gesture, slide with annotation or by sending some link through chat. If necessary, he can draw some figure on the whiteboard panel to improve the students' understanding as shown in the Fig. 4.

Currently, we are working on a prototype of the system. About the development platform, the client apps for instructor and student on the tablet are being implemented in Swift language [6] on iOS [7] using Xcode [8] integrated development environment on macOS Mojave. The tablet used in our system is 9.7 in. iPad (6th generation) with iOS 12. For the server development, C++ language is being used with GNU compiler on Linux.

3 Conclusion

We proposed a design of a tablet-based live learning system with whiteboard support to improve annotation capability. In this system, when an instructor gives a lecture in front of his iPad, students watch the lecture with presentation slide with annotation in the slide page on their iPad in real time. They can also ask questions using a chat panel. However, since there is not enough space for drawing large figure for detailed annotation by the instructor, we improved annotation capability by supporting a whiteboard component in the user interface of the iPad client. We believe that this whiteboard support will help give students deeper understanding of the lecture on their iPad in real time.

Currently, we are working on the prototype implementation. When it is finished, we plan to have a group of students in our department test it and then to survey on how much the whiteboard capability of our tablet-based live learning system help them understand the lecture in real time.

Acknowledgements This work was supported by 2017 Hongik University Research Fund.

References

1. Wains SI, Mahmood W (2008) Integrating M-learning with E-learning. In: 9th ACM SIGITE conference on information technology education. ACM, pp 31–38
2. Anderson R, Anderson R, Davis P, Linnell N, Prince C Razmov V, Videon F (2007) Classroom presenter: enhancing interactive education with digital ink. *IEEE Comput* 40(9):56–61
3. Ulrich C, Shen R, Tong R, Tan X (2010) A mobile live video learning system for large-scale learning-system design and evaluation. *IEEE Trans Learn Technol* 3(1):6–17

4. Lee J (2015) Live mobile learning system with enhanced user interaction. In: Park DS et al (eds) *Advances in computer science and ubiquitous computing (CUTE 2015)*. LNEE, vol 373. Springer, pp 745–750
5. HTTP Live Streaming. https://developer.apple.com/documentation/http_live_streaming
6. Swift 5.1. <https://swift.org>
7. iOS 13. <https://developer.apple.com/ios/>
8. Xcode 11. <https://developer.apple.com/xcode/>

Gearbox Fault Diagnosis Under Variable Speed Condition Using Frequency Spectral Analysis with 1D Residual Neural Network



Md Arafat Habib and Jong-Myon Kim

Abstract Stringent classification of gearbox fault conditions is indispensable for industrial safety. Determining the best set of features by analyzing the statistical parameters of the signals is one of the most climacteric tasks in data-driven fault diagnosis. For variable speed conditions, variant fault types of gearboxes have dynamic characteristics and these statistical features fail to unveil them. To address this issue, deep learning algorithms are used to produce a better performance of the feature selection process. In this paper, a combination of frequency spectral analysis of the acoustic emission signals and a 1-dimensional residual neural network (1D-RNN) is proposed. Our proposed method through the processed 1D-RNN shows vigorous classification performance, resulting in up to 95.6% classification accuracy in all the considered scenarios.

Keywords Gearbox safety · Fault diagnosis · Convolutional neural network

1 Introduction

An effective fault diagnosis approach for gearboxes under invariant speed conditions can reduce protection expenditure and ensure function dependability. Research based on data-driven fault diagnosis can reduce the costs of preservation through the consistency of the machinery [1]. Acoustic emission (AE) signals can extract intrinsic information from low energy signals. AE signals can be more effective for data-driven fault diagnosis than vibration signals [2]. Different approaches have been proposed [3, 4] to prove the domain-based classification by analyzing the extracted features from signals. Unfortunately, these approaches suffer from the inappropriate time-window adjustment and are unable to capture high frequency resolution.

M. A. Habib · J.-M. Kim (✉)

Department of Electrical/Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, Republic of Korea
e-mail: jmkim07@ulsan.ac.kr

M. A. Habib

e-mail: akhtab007@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_31

227

The proposed method in this paper first takes the raw input signal and denoises it. After that, to get the positive frequency responses, a fast Fourier transformation (FFT) for the denoised signals is calculated. For the final fault condition analysis, a 1-dimensional residual neural network (1D-RNN) was used. To establish the superiority of our proposed approach, we compared it with various state-of-the-art algorithms such as neural networks using statistical parameters accompanied by a multi-class support vector machine [5], as well as the spectral average with the k-nearest neighbor algorithm (KNN) [6].

The major contributions of this work can be summarized as follows:

- The frequency domain knowledge has been used instead of the statistical features at different speeds for gearboxes to analyze the effectiveness of the AE signals.
- Frequency responses collected using FFT have been used as an input to the proposed 1D-RNN for fault classification in a speed-invariant way. To validate our method, we also conducted experiments under different speed conditions (revolutions per minute [RPMs]) and compared them with existing state-of-the-art algorithms.

The rest of the paper is organized as follows. In the following section, details of the proposed methodology are presented, including the gearbox data acquisition testbed. In Sect. 3, experimental results are presented to establish the robustness of the proposed method. Finally, this paper is concluded in Sect. 4.

2 Methodology

There are three major sections of the proposed method. The data are collected from an experimental testbed. The FFT is calculated to get the positive frequency, and the classification using 1D-RNN follows this. AE sensors were used in the experimental setup. Signals are collected from the sensors of the bearing housing end from two channels. AE signals after passing through FFT are used as input for 1D-RNN.

2.1 Data Acquisition

For our experimental purpose, we considered a simple gearbox with a gear ratio of 1.52:1. Two shafts, a non-drive-end shaft (NDS) and a drive-end shaft (DS), are connected in the experimental setup. At three different RPMs (300, 600, and 900), a three-phase induction motor is connected with a displacement transducer. Through the gearbox, the bearing house is attached to the motor shaft. A WS α AE sensor is kept over the bearing house in the shaft at the NDS. Using a PCI-2 system, AE signals are collected through the AE sensor. The signals are collected at a sampling rate of 100,000 Hz. A real-time test bed scenario for the data collection is depicted in Fig. 1a. Figure 1b presents the experimental testbed in detail. The specification

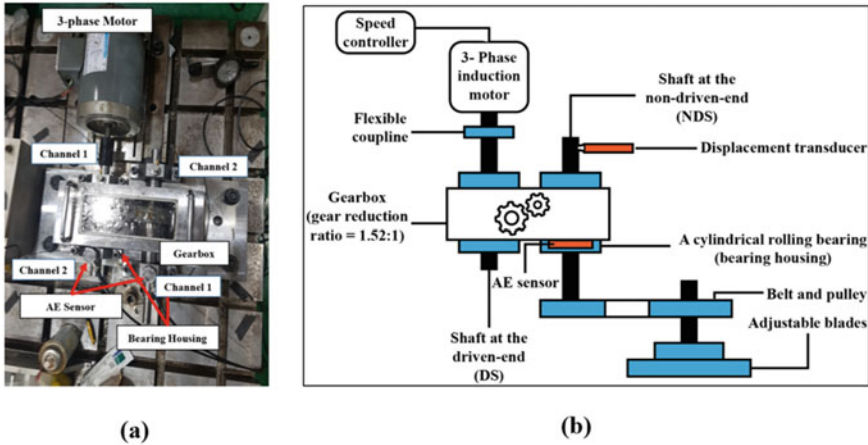


Fig. 1 a Real-time testbed for the experiment. b Schematic of the experimental testbed for gearbox fault identification

Table 1 Detailed gear specification

Gearbox specification	Value
Number of draft shaft teeth	25
Number of driven coaxial teeth	38
Tooth length	9

Table 2 Specifications of the defective coaxial driven shaft gear

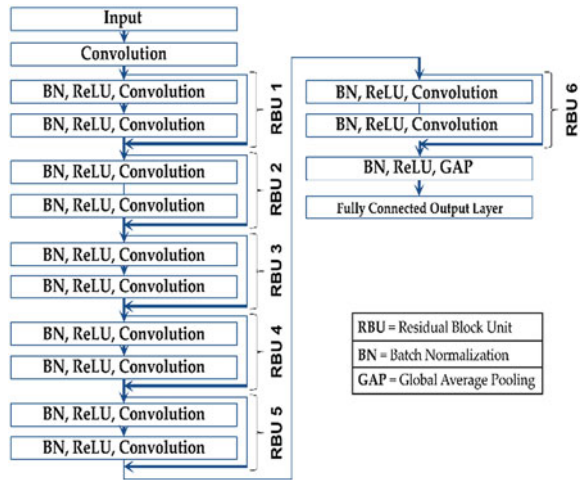
Health condition	Defect length (mm)
10% crack (C10)	0.9
20% crack (C20)	1.8
30% crack (C30)	2.7

of the gears used in our setup is described in Table 1. An artificial defect is created on the shaft gear to create variant health conditions. The specifications of the faulty health conditions of the gear are described in Table 2.

2.2 Calculating FFT to Get the Positive Frequency Data Acquisition and 1D-RNN

The unwanted noise in the raw AE signal after collection is removed using a white-noise cancellation process. FFT is used next to accumulate the positive frequency response from the input signal. The AE spectrum consists of 5×10^4 positive

Fig. 2 Block diagram of the proposed 1-dimensional residual neural network



frequency components. This is not a suitable input for 1D-RNN, which is why they are divided into several frequency bins.

A residual network has convolution layers, fully connected layers, activation functions, batch normalizations, and a feed-forward nature [7, 8] like CNN. The only difference a residual network has is the additional residual block. In our proposed 1D-RNN, 6 residual energy blocks were used. For intermediate layers, a ReLU (rectified liner unit) activation layer was used, and for the output layer, a SoftMax classifier was used. For the optimizer, Adam was used. Figure 2 illustrates the entire architecture.

3 Results Analysis and Discussion

To assess the performance of the proposed 1D-RNN fault classification approach, the following methods were used.

- 1D residual neural-network-based invariant gearbox health state visualization.
- RPM-invariant performance analysis of health state classification using [5] and [6].

Specifications of the collected dataset are given in Table 3.

As discussed in Table 3, we considered two different datasets for the validation of our proposed approach. The datasets have different speeds with similar health conditions. In our first scenario, dataset 1 is used for training and dataset 2 is used for the classification test. In the second scenario, dataset 2 is used for training and data set 1 is used for classification. We have considered three performance metrics: the F1 score, average classification accuracy (AC), and overall classification accuracy (OC) [9].

Table 3 Details of the considered working conditions with the same health types

	Health type	Shaft speed (rpm)	Sampling frequency (Hz)
Dataset 1	Normal Condition (NC)	300	100,0000
	10% Crack (C10)	300	
	20% Crack (C20)	300	
	30% Crack (C30)	300	
Dataset 2	Normal Condition (NC)	900	
	10% Crack (C10)	900	
	20% Crack (C20)	900	
	30% Crack (C30)	900	

The F1 score is very important for balancing between the recall and precision scores. It can be calculated using the following equation:

$$F1 = \frac{T_p}{T_p + (F_n + F_p)/2} \times 100\% \tag{1}$$

where T_p is the number of correctly classified samples from a particular class and F_n is the number of incorrectly classified samples from a particular class. The final result is obtained as a percentage. After calculating the F1 score, the average accuracy is calculated as follows:

$$AC = \frac{\sum F1}{\sum T_c} \tag{2}$$

where T_c is the total number of classes. Finally, the overall classification accuracy (OC) is calculated by considering the average of all ACs from different scenarios. In this experiment, eightfold cross validation is used.

The performance analysis of the proposed method is given in Table 4. It achieves 95.6% overall accuracy. To prove the superiority of the proposed method, we show a comparative analysis with existing state-of-the art algorithms. For a comparative analysis with the proposed 1D-RNN, the following have been implemented with the same data set: a multiclass support vector machine with a neural net that considers statistical features [5], and the spectral average accompanied by KNN [6]. A comparison of the classification accuracy is presented in Table 5. Our proposed method

Table 4 Analytical implementation of the proposed model for various scenarios

Scenario	Training Dataset	Test Dataset	F1 (%)				CA (%)	Overall
			NC	C10	C20	C30		
1	Dataset 1	Dataset 2	94.96	93.42	95.49	95.59	94.87	95.6
2	Dataset 2	Dataset 1	95.44	95.37	96.79	95.42	96.33	

Table 5 Comparative analysis of different methods

Scenario	Method	F1 (%)				AC (%)	Improved (%)
		NC	C10	C20	C30		
1	Jin et al. [5]	87.52	89.4	88.9	91.4	89.31	5.56
	Yoon et al. [6]	45.21	48.52	47.22	47.9	47.21	47.66
	Proposed	94.96	93.42	95.49	95.59	94.87	–
2	Jin et al. [5]	87.93	89.91	87.2	88.6	88.41	7.92
	Yoon et al. [6]	47.27	49.11	48.73	47.44	48.14	48.19
	Proposed	95.44	95.37	96.79	95.42	96.33	–

outsmarts the methods in [5] and [6] by at least 5.56% accuracy for each scenario. Identical settings for training and testing the data have been used.

4 Conclusions

This paper proposed a 1-dimensional residual neural-network-based classifier that uses frequency domain features for fault classification of gearboxes. The proposed method is invariant to the shaft speed. Many of the previous approaches opt for statistical features that are not always robust for gearbox classification. This study considers an invariant scenario for different fault conditions and incorporates different RPMs (300 and 900). The proposed 1D-RNN achieved an overall classification accuracy of 95.6%. Additionally, the proposed approach outperformed state-of-the-art approaches with an improvement of at least 5.75%.

Acknowledgements This research was financially supported by the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea and Korea Institute for Advancement of Technology (KIAT) through the Encouragement Program for The Industries of Economic Cooperation Region (P0006123).

References

1. Sohaib M, Kim C, Kim J (2017) A hybrid feature model and deep-learning-based bearing fault diagnosis. *Sensors*. 17:2876
2. Eftekharnjad B, Carrasco M, Charnley B, Mba D (2011) The application of spectral kurtosis on acoustic emission and vibrations from a defective bearing. *Mech Syst Signal Process*. 25:266–284
3. Kim B, Lee S, Lee M, Ni J, Song J, Lee C (2007) A comparative study on damage detection in speed-up and coast-down process of grinding spindle-typed rotor-bearing system. *J Mater Process Technol* 187–188:30–36
4. Tahir M, Khan A, Iqbal N, Hussain A, Badshah S (2017) Enhancing fault classification accuracy of ball bearing using central tendency based time domain features. *IEEE Access*. 5:72–83
5. Jin X, Zhao M, Chow T, Pecht M (2014) Motor bearing fault diagnosis using trace ratio linear discriminant analysis. *IEEE Trans Industr Electron* 61:2441–2451
6. Yoon J, He D, Van Hecke B, Nostrand T, Zhu J, Bechhoefer E (2015) Vibration-based wind turbine planetary gearbox fault diagnosis using spectral averaging. *Wind Energy*. 19:1733–1747
7. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86:2278–2324
8. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444
9. Hasan M, Islam M, Kim J (2019) Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. *Measurement* 138:620–631

Health State Classification of a Spherical Tank Using a Hybrid Bag of Features and k-Nearest Neighbor



Md Junayed Hasan, Jaeyoung Kim, and Jong-Myon Kim

Abstract Feature analysis plays an important role in determining the various health conditions of mechanical vessels. To achieve balance between traditional feature extraction and the automated feature selection process, a hybrid bag of features (HBoF) is designed for the health state classification of spherical tanks in this paper. The proposed HBoF is composed of (a) the acoustic emission (AE) features, and (b) the time and frequency based statistical features. A wrapper-based feature selector algorithm, Boruta, is applied to extract the most intrinsic feature set from HBoF. The selective feature matrix is passed to the k-nearest neighbor (k-NN) classifier to distinguish between normal condition (NC) and faulty condition (FC). Experimental results show that the proposed approach yields an average 100% accuracy for all working conditions. The proposed method outperforms the existing state-of-the-art approaches by achieving at least 19% higher classification accuracy.

Keywords Spherical tank · AE features · Boruta · Fault diagnosis

1 Introduction

Mechanical vessels play a very important role in day-to-day life with widespread application [1]. Specifically, in the oil and gas industry, the use of spherical tanks is required due to the cost effectiveness of building a sphere. With the increasing use of these types of spherical tanks for different industries, the number of accidents related to leakage from the bottoms of these tanks is also increasing [2]. As a result, improved safety precautions and maintenance are required [3, 4].

M. J. Hasan · J. Kim · J.-M. Kim (✉)

Department of Electrical and Computer Engineering, University of Ulsan, Ulsan, South Korea

e-mail: jongmyon.kim@gmail.com

M. J. Hasan

e-mail: junhasan@gmail.com

J. Kim

e-mail: kjy7097@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_32

In this experiment, the main emphasis is on health state categorization through signals acquired from a spherical tank. Identifying the health condition (normal or faulty state) through signals at an early stage will make it easier to determine the necessary precautions at later stages. The acoustic emission (AE) velocity signals are considered for classification of the health state. Compared with old-style methods, AE is an economical and efficient detection process [5]. Additionally, AE signals can provide underlying information from low energy signals [6, 7] for a more substantial data-driven fault identification approach. AE-based diagnosis methods mostly rely on a procedure for analyzing the peak of the characteristic frequencies of the signals [8]. Pattern generation from acquired signal domains using several signal-imaging techniques can also differentiate between health conditions for further classification [9]. Several automated feature learning processes driven by deep learning-based algorithms have been studied to reduce the necessity of domain knowledge expertise [9–11]. Due to limitations in the amount of data, deep learning-based approaches are not capable of extracting meaningful features.

Herein, a data-driven hybrid feature extraction process is considered. The main contributions of this research can be summarized as follows. (1) An HBoF extraction method is designed by combining two types of analysis: analysis of the AE signal properties, and of the time-domain and frequency-domain based statistical properties; and (2) a wrapper-based non-redundant feature selection method, Boruta, is utilized to analyze all the key elements of the hybrid feature pool. Finally, the k-nearest neighborhood (k-NN) is applied for classification of the health state, using those selected features as input.

The rest of the paper is structured as follows. Section 2 provides details of the methodology, including the AE data acquisition system. The analysis of the experimental results and comparative discussions are provided in Sect. 3. The paper is concluded in Sect. 4.

2 Proposed Method

The proposed approach is divided into four sections: (1) data collection from a multisensory testbed, (2) feature extraction by HBoF, (3) feature selection by Boruta, and (4) k-NN-based classification.

2.1 *Experimental Testbed and Dataset Acquisition*

An experiment is performed on a self-designed test platform to collect AE signals. One AE sensor (WDI-AST) with four different channels is attached to collect the velocity AE signals. On the test rig, there are four different crack positions (825, 750, 1040 and 430 mm) to collect the velocity data with 1 MHz sampling frequency.

The signal is measured through a trigger-based measurement technique for a specific amount of time.

2.2 Hybrid Bag of Features

It is difficult to obtain intrinsic information for different health types from a raw signal. To create the health condition-based feature matrix, two different sets of features are considered. For the AE features, the amplitude (F1), rise time (F2), and duration (F3) of the signals are computed. For the threshold value, the rms of the signal is considered. The specifics of the AE features are demonstrated in Fig. 1. For statistical analysis, from the time domain, the numerical features obtained are root mean square (F4), kurtosis (F5), skewness (F6), shape factor (F7), and impulse factor (F8). In the same manner, from the frequency domain, the features obtained are root mean square (F9), kurtosis (F10), and skewness (F11). Thus, in total, 11 features are extracted to create the designed HBoF. In Table 1, the numerical details of these statistical features are described.

Fig. 1 Illustration of acoustic emission (AE) signal feature

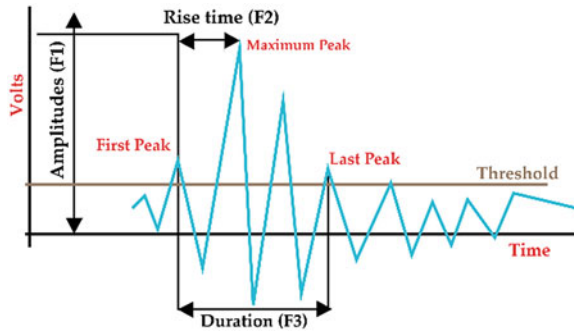


Table 1 Numerical explanation of statistical features

Feature	Equation	Feature	Equation	Feature	Equation
F4	$\sqrt{\frac{1}{N} \sum_{i=1}^N X_i^2}$	F5	$\frac{1}{N} \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{\sigma} \right)^4$	F6	$\frac{1}{N} \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{\sigma} \right)^3$
F7	$\frac{\frac{1}{N} \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{\sigma} \right)^3}{\frac{1}{N} \sum_{i=1}^N X_i }$	F8	$\frac{\max(X)}{\frac{1}{N} \sum_{i=1}^N X_i }$		
F9	$\sqrt{\frac{1}{N} \sum_{i=1}^N F_i^2}$	F10	$\frac{1}{N} \sum_{i=1}^N \left(\frac{F_i - \bar{F}}{\sigma} \right)^4$	F11	$\frac{1}{N} \sum_{i=1}^N \left(\frac{F_i - \bar{F}}{\sigma} \right)^3$

Here, x is the time-domain raw signal, and F is the frequency domain signal. N is the total number of samples

2.3 Feature Selection by Boruta

Boruta finds the most relevant and intrinsic feature information from data. As a first step, it duplicates the original feature set and then rearranges the feature values, which are called shadow features. Each of those shadow feature sub-sets is then trained by the random forest classifier to validate the significance of the important feature set by the mean decrease impurity (MDI) matrix. If the MDI value is higher, then the set is important. In the second step, it runs a similar test for the original feature set. For this test, Z score is calculated. Z score is the number of standard deviations a measure is from the mean. The algorithm considers whether the original set of features has a higher Z score than most of its shadow features. If the score is high, it is logged as a vector named hits. Thus, the iteration is continued till reaching the predefined set of iteration numbers and, at the end, a hit table is generated.

2.4 Fault Classification Using *k*-Nearest Neighbor (*k*-NN)

To validate the considered optimal feature sets in terms of classification performance, a *k*-NN classifier is used. *k*-NN has a simple architecture with less computational complexity [6]. *k*-NN categorizes the trials relying on the votes of the *k*-nearest neighbors, which are identified by certain distance parameters [12].

3 Experimental Result Analysis and Discussion

3.1 Dataset

The standard AE dataset of spherical tanks is used to conduct a test. A 0.1 s velocity signal with 1 MHz sampling frequency is used for consideration of each health state (NC and FC). The particulars of the dataset are provided in Table 2.

Table 2 Details of the considered dataset

Health condition	Crack type	Crack size (mm)	Channels
Normal condition (NC)	No crack	No crack	4
Faulty condition (FC)	Pinhole crack	3	4

3.2 Result Analysis

Raw AE signals have no intrinsic information to reveal different health conditions. Therefore, the HBoF is designed and Boruta is applied to get the most intrinsic feature information. From Boruta, the five most important features are calculated (i.e., F1, F3, F4, F5, and F10). These five features are collected from AE analysis, time domain, and frequency domain. The robustness of the HBoF is shown by considering all the important information from the signals.

The selected features from Boruta are each provided to the k-NN. The dataset is divided into training and testing sets at the respective proportion of 60/40. Sensitivity is considered for calculating the classwise accuracy. The final classification accuracy is obtained after 6-fold cross validation. The proposed approach achieves 100% classification accuracy when the optimal value of k is 8 in k-NN algorithm (illustrated in Fig. 2b). Along with the proposed approach, several comparisons are made to establish the robustness. From the HBoF, for feature selection, instead of Boruta, non-dimensional feature reduction techniques such as PCA and t-SNE are applied to get the intrinsic feature information for final classification. In Table 3, the classification

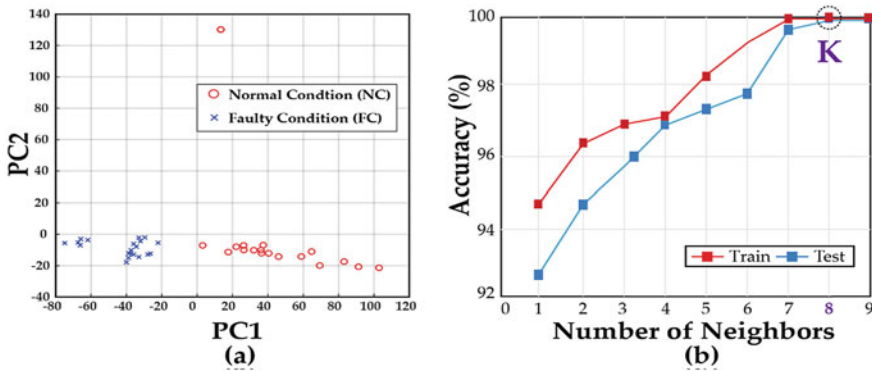


Fig. 2 a Boruta feature space. The five features selected from Boruta are embedded into 2D space by PCA for visualization purposes only. b Various categorization accuracies as a function of the number of neighbors (k). The optimal value is k = 8

Table 3 Classification accuracy of different methods

Approach	Classification accuracy (%)		Average classification accuracy (%)	Decrement from the proposed method (%)
	NC	FC		
Proposed	100	100	100	–
HBoF + t-SNE + k-NN	75.5	35	55.25	44.75
HBoF + PCA + k-NN	79.5	82.5	81	19

accuracies from different approaches are described in a very detailed way. From the information shown there, the necessity of finding out the ranked features is demonstrated, as opposed to keeping all the information.

4 Conclusion

This paper presented a hybrid feature selection method called HBoF, which is composed of AE feature analysis and statistical information from time and frequency analysis. To select the most intrinsic features from the proposed HBoF, feature wrapper Boruta is applied. Thereafter, k-NN is used for final classification, which leads to a 100% average accuracy for both normal and faulty conditions (NC and FC). Comparative analysis with different non-linear feature dimensionality reduction techniques (i.e.; PCA, and t-SNE) was performed to validate the performance. The proposed approach outperformed the PCA and t-SNE based methods by respective 19% and 44.75% classification accuracies.

Acknowledgements This work was supported by the Technology Infrastructure Program funded by the Ministry of SMEs and Startups (MSS, Korea).

References

1. Saidur R (2010) A review on electrical motors energy use and energy savings. *Renew Sustain Energy Rev* 14:877–898
2. Morofuji K, Tsui N, Yamada M, Maie A, Yuyama S, Li ZW (2003) Quantitative study of acoustic emission due to leaks from water tanks. *Group*. 21:213–222
3. Luo T, Wu C, Duan L (2018) Fishbone diagram and risk matrix analysis method and its application in safety assessment of natural gas spherical tank. *J Clean Prod* 174:296–304
4. Korkmaz KA, Sari A, Carhoglu AI (2011) Seismic risk assessment of storage tanks in Turkish industrial facilities. *J Loss Prev Process Ind* 24:314–320
5. Li W, Dai G, Wang Y, Long F (2011) Study of tank acoustic emission testing signals analysis method based on wavelet neural network. In: *ASME pressure vessels and piping conference*
6. Pandya DH, Upadhyay SH, Harsha SP (2013) Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. *Expert Syst Appl* 40:4137–4145
7. Niknam SA, Songmene V, Au YHJ (2013) The use of acoustic emission information to distinguish between dry and lubricated rolling element bearings in low-speed rotating machines. *Int J Adv Manuf Technol* 69:2679–2689
8. Kang M, Kim J, Kim J (2015) High-performance and energy-efficient fault diagnosis using effective envelope analysis processing unit. *IEEE Trans Power Electron* 30:2763–2776
9. Amar M, Gondal I, Wilson C (2015) Vibration spectrum imaging: a novel bearing fault classification approach. *IEEE Trans Ind Electron* 62:494–502
10. Sohaib M, Kim C-H, Kim J-M (2017) A hybrid feature model and deep-learning-based bearing fault diagnosis. *Sensors* 17:2876
11. Hasan MJ, Sohaib M, Kim J-M (2019) 1D CNN-based transfer learning model for bearing fault diagnosis under variable working conditions

12. Chen X, Xu J-B, Guo W-Q (2013) The research about video surveillance platform based on cloud computing. In: 2013 international conference on machine learning and cybernetics. IEEE, pp 979–983

L-RDF Diversity: Distributed De-Identification for Large RDF Data with Spark



Minhyuk Jeon, Odsuren Temuujin, Yoonmi Shin, Jinhyun Ahn,
and Dong-Hyuk Im

Abstract Privacy protection issues for RDFs (resource description framework) have emerged with the using of public government open data, healthcare data for individuals. As that data may include personal information, it must go through a de-identification process that deletes or replaces part of the original data. To enable these protections, a method has been developed to apply k-anonymization for RDF data. However, sensitive RDF information anonymized using k-anonymization is not entirely secure and is vulnerable to attacks. In this paper, we use an l-diversity Anatomy de-identification method that can overcome the limitations of k-anonymity and guarantee stronger privacy protection than k-anonymization. Since this process for anonymization of data requires a lot of computational time, we use Spark distribution computing to provide rapid de-identification to enhance its utility.

Keywords Privacy protection · RDF · De-identification · l-diversity · Anatomy · Spark

M. Jeon · O. Temuujin · Y. Shin · D.-H. Im (✉)
Hoseo University, Asan-si, Chungcheongnam-do 31499, Korea
e-mail: dhim@hoseo.edu

M. Jeon
e-mail: jeoncoder@gmail.com

O. Temuujin
e-mail: temuujintemka@gmail.com

Y. Shin
e-mail: sinyoonmi12@gmail.com

J. Ahn
Jeju National University, Jeju 63243, Republic of Korea
e-mail: jha@hoseo.edu

1 Introduction

Increasing volumes of RDF (resource description framework) data are being created and exchanged on the web, which often includes transfer of private information. To prevent the leakage of such information, we need a de-identification process that removes the identity of the individuals associated with the data. So that, several models have been researched for de-identification, such as k-anonymity [1], l-diversity [2] model. However, the application of these models for RDF data is still the only observed use of k-anonymity [3, 4].

The l-diversity model requires at least l distinct value for a sensitive attribute, and this model also includes the benefits of k-anonymity. And the Anatomy algorithm [5] is the most widely used privacy protection algorithm for the l-diversity model, because it can achieve de-identification without using generalization or suppression. However, it is difficult to apply the Anatomy algorithm for the de-identification of RDF data that have different structures. In addition, this anonymization process can take a long time to apply to large-scale RDF data. Thus, we develop a de-identification algorithm for large-scale RDF data with a big data processing platform such as the Apache Spark distributed processing platform [6]. To the best of our knowledge, ours is the first study devoted to applying the Anatomy algorithm for RDF anonymization model in Spark.

2 The Proposed Approach

In this section, we applied a de-identification process algorithm to the RDF model. First, we parse and set input data from RDF data to property table. This transformation is shown in Fig. 1.

Next we set the l value to 2. Then we eliminate (ID) identifiers such as social security numbers, that allow you to directly specify a particular person. Attributes that can be identified by combining them with at least two other attributes are classified as quasi-identifier (QI). We also select a sensitive attribute (SA) that is the most sensitive attribute that can be targeted by an attack. In the example shown in Fig. 2, the ID is "Name" and "SSN", QI is "Gender" and "Address", and SA is "Disease".

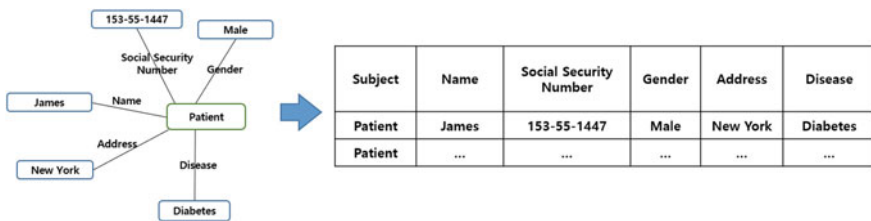


Fig. 1 RDF data and property table

Subject	Name	Social Security Number	Gender	Address	Disease
Patient	James	55-55-1447	Male	New York	Diabetes
Patient	Frank	352-71-7134	Female	Boston	Cancer
Patient	Boil	442-11-5602	Male	Washington	Aids
Patient	Smith	134-11-2319	Male	New Jersey	Aids
Patient	Jane	140-12-8841	Female	New York	Osteoporosis

ID

QI

SA

Fig. 2 Depiction of how the useful attributes remain

As shown in Fig. 3’s above image, this classified data is divided into SABuckets with each SA as a key. We also created a new group. Next, the data is moved to this group one by one for each bucket. If the group contains more than 1 data, it creates the next group and repeats the same task. Finally, less than 1 data may remain, in which case they are added to a random group.

In Fig. 3’s below image, a group of such QIs is called an equivalence class (EC). All of these ECs and group numbers are combined into one table as a QI table (QIT), and the group number, SA, and the number of SAs present in the group are combined as an SA table (ST).

In the generated QIT and ST, only the IDs are lost for existing RDF data. And un-like other de-Identification models, the QI is not generalized. Therefore, it is more valuable for research or statistics. These tables do not identify any specific individual, and there are two distinct SAs in Group 1 and three in Group 2. Therefore, the condition of the l-diversity model is satisfied. This QIT and ST are shared in the form of a blank node in Turtle format, so that each group can be identified and referenced.

And configuring the Anatomy algorithm with Spark requires an additional data type called Resilient Disributed Data (RDD), although its basic framework is similar to that of the existing Anatomy algorithm. The pseudocode of the total algorithm using RDD is shown in Fig. 4.

3 Experimental Evaluation

In this experiment, groups of 10, 50, 100 and 300 universities were used, in Turtle format. And we used a cluster system consisting of four machines, and each machine has an Intel (R) Xeon (R) CPU E3-1220 V2 @ 3.10 GHz CPU, with a memory of 24 Gb. Also, for the experiments, we modified the synthetic data. For example, we only used professor types such as {fullProfessor, associateProfessor, assistantProfessor} in-formation for de-identification. Explicit identifiers of professors were deleted from the data, and the properties {name, researchInterest, undergraduateDegreeFrom, masterDegreeFrom, doctorDegreeFrom} were used.

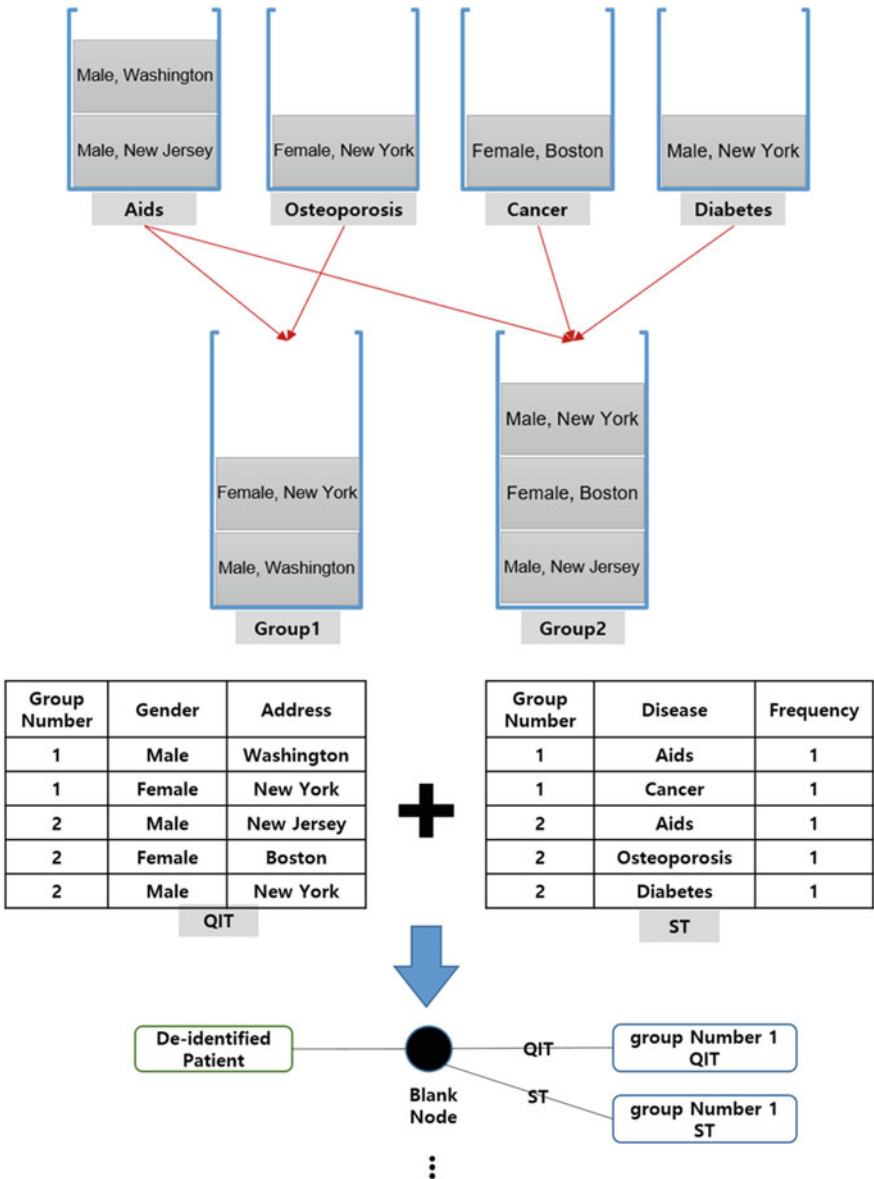


Fig. 3 Creation of a QIT and ST (1 value = 2)

Data: HDFS RDF file, L value
Result: De-identified RDF file

```

// Lines 1-6 are the Conversion stages of RDF to tuples.
1. RDF is divided into as many partitions as any number at least
   executor's count.
2. tripleList = The List of associate Triples.
3. tripleRDD = parallelized tripleList.
4. mapTripleRDD = Mapping tripleRDD as {SA, SA's Tuple}.
5. SASet = mapTripleRDD's SA keySet.
6. SABucket; bucketCnt = 0; groupCnt = 0.
// Lines 7-8 are the Creation stage of the buckets, which are filled with
SAs.
7. For each loopSA in SASet
8.     SABucketbucketCnt = Add a list by filtering so that the keys in
   mapTripleRDD are the same as loopSA.
9.     bucketCnt = bucketCnt + 1.
// Lines 10-19 are Creation stages of groups which are filled with Tuples.
10. While If there are at least L non-empty SABucket
11.     Sort an SABucket by SABucket member's Count.
12.     new groupBucket
13.     for idx=1 to L
14.         tuple = Get a value in SABucket's idx index.
15.         Add tuple's first value and groupCnt to the groupBucket.
16.         remove a value which is added to groupBucket in SABucket.
17.         groupCnt = groupCnt + 1.
18. For each bucket in non-empty SABuckets.
19.     Add bucket's first value and a random number within the
   groupCnt range to the groupBucket.
// Lines 20-22 are Division stages to QIT & ST.
20. allTuples = parallelised groupBucket.
21. QIT = Extract the required attributes from Tuples to create a QI
   partition and coalesce all partitions.
22. ST = Extract the required attributes from Tuples to create a SA
   partition and coalesce all partitions.
// Lines 23-24 are Creation stages of RDF consisting of QIT & ST.
23. For each cnt in groupCnt
24.     Add a QITcnt, SAcnt Triple through Jena.

```

Fig. 4 Spark-based algorithm for RDF

So, we proceeded with two methods: de-identification on the Java InMemory, and an RDD code on Java Spark distributed clusters. First, In the case when the size of input data is small, the InMemory operation ran faster than Spark. However, as LUBM increased in size, for InMemory, the computational speed de-created. Moreover, data above LUBM-(300,0) could not be executed due to lack of hardware resources. Spark

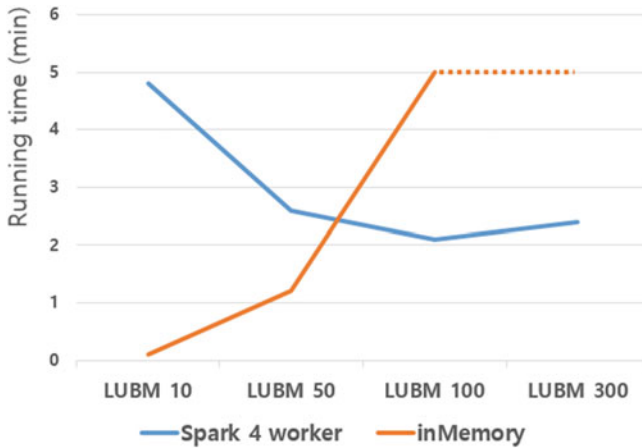


Fig. 5 Graph showing performance comparison

consists of a manager that runs the driver and an executor that performs the actual operation. In this experiment, we used Spark-submit to help Spark run, giving us several optimal options, including the number of CPU cores and amount of memory to use.

In Fig. 5, graph compares InMemory and Spark, which uses 4 workers. As the graph shows, InMemory is more powerful for small data, but it cannot support big files. Spark does not show stable results with low volumes of data. And, if the substitution of a worker that is better than the cluster specification used in the current experiment is done, faster de-identification for larger RDF data can be supported.

4 Conclusion and Future Work

In this study, we proposed a platform for distributing large-scale RDF data on Spark and de-identifying it using the Anatomy algorithm that satisfies l -diversity. An experimental evaluation demonstrated that a Spark-based Anatomy algorithm demonstrated a significant advantage with large RDF datasets.

Although the Anatomy algorithm is applied with the goal of data preservation, it is still vulnerable to inference attack. To solve this problem, it is necessary to apply an algorithm that satisfies the t -proximity model and preserves the utility of the data like the Anatomy algorithm does. In addition, we will focus on l -diversity algorithms for dynamically published datasets [7].

Acknowledgements This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant NRF-2017R1C1B1003600, in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program Supervised by the Institute

for Information & Communications Technology Promotion (IITP), under Grant IITP-2019-2018-0-01417, and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A1B07048380.

References

1. Sweeney L (2002) k-anonymity: a model for protecting privacy. *Int J Uncertain, Fuzziness Knowl-Based Syst* 10(05):557–570
2. Machanavajjhala A et al (2006) l-diversity: privacy beyond k-anonymity. In: 22nd international conference on data engineering (ICDE'06). IEEE, pp 24–24
3. Radulovic F, García Castro, R, Gómez-Pérez A (2015) Towards the anonymisation of RDF data
4. Temuujin O et al (2019) SPARK-based partitioning algorithm for k-anonymization of large RDFs. In: *Advanced multimedia and ubiquitous engineering*. Springer, Singapore, pp 292–298
5. Xiao X, Tao Y (2006) Anatomy: simple and effective privacy preservation. In: *Pro-ceedings of the 32nd international conference on very large data bases*. VLDB Endowment, pp 139–150
6. Zaharia M et al (2016) Apache spark: a unified engine for big data processing. *Communications of the ACM* 59(11):56–65
7. Temuujin O, Ahn J, Im D-H (2019) Efficient L-diversity algorithm for preserving privacy of dynamically published datasets. *IEEE Access* 7:122878–122888

Intelligent Personalized Transport Alert System with Edge Computing



Hyolin Choi, Jiwon Hong, and Yongik Yoon

Abstract Most people's lives are based on a repetitive routine. Despite this fact, people check information on traffic situations manually using their mobile applications every day. Also, inconveniences may be caused due to unexpected and constantly changing traffic situations that arise from a number of factors. Many services that provide information on traffic, weather, and transportation are available, but there isn't a system that provides all this information at the same time. As a solution to this problem, we suggest a new notion called Intelligent Personalized Transport Alert System (IPTAS). IPTAS provides information on the transportation mode and the arrival time to users automatically via speech and text notification based on the user's current location, time, date, weather, and traffic situation. Through IPTAS, convenience in daily life is enhanced, and more accurate information is given to the user.

Keywords Artificial intelligence system · Personalized system · Public data · Real-time android · Notification alert system

1 Introduction

One of the many inconveniences people face on a daily basis is checking for transportation information and waiting for the arrival of their transportation. The waiting time is affected by traffic situations. There are many factors that attribute to the

H. Choi (✉) · J. Hong · Y. Yoon
Department of Information Technology Engineering, Sookmyung Women's University, Seoul,
South Korea
e-mail: choehyolin@gmail.com

J. Hong
e-mail: jiwon_h98@naver.com

Y. Yoon
e-mail: yiyoony@sookmyung.ac.kr

traffic situation such as weather and time, and because of this, it is difficult for users to decide the most ideal transportation mode for themselves.

This paper researches on a system that automatically provides transportation information via speech and text notification based on the analysis of the user's route, transportation mode, step count, current weather, and traffic situation. The result of this research enhances convenience by reducing the need for users to check for transportation information in person, and by providing notification alert automatically.

The purpose of Intelligent Personalized Transport Alert System is providing speech and text notification on transportation information that matches with the user's machine-learned route based on the user's current location, date, and time, and recommending whether to take a bus/subway or to walk based on the user's step count when the travel distance is less than 3 km.

The plan of this paper is as follows: the second part consists of the flow diagram and the architecture, the third part consists of the system's implementation environment and the results, and the fourth part includes the conclusion of this paper.

2 Intelligent Model with Edge Computing

Figure 1 shows the flow diagram of Intelligent Personalized Transport Alert System. The user's mobile device generates the data on the day of the week, weather, and traffic status using offloading policy. Pattern Knowledge Device then matches time and location with the data in the database to provide transportation information based

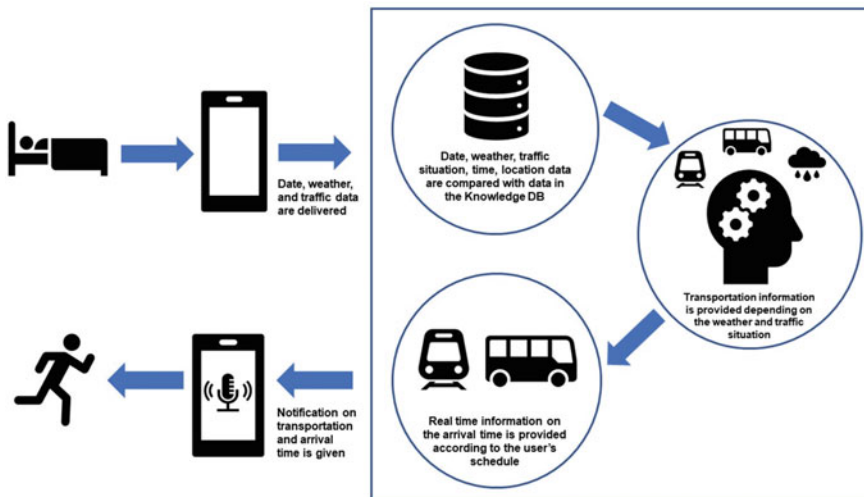


Fig. 1 System flow diagram

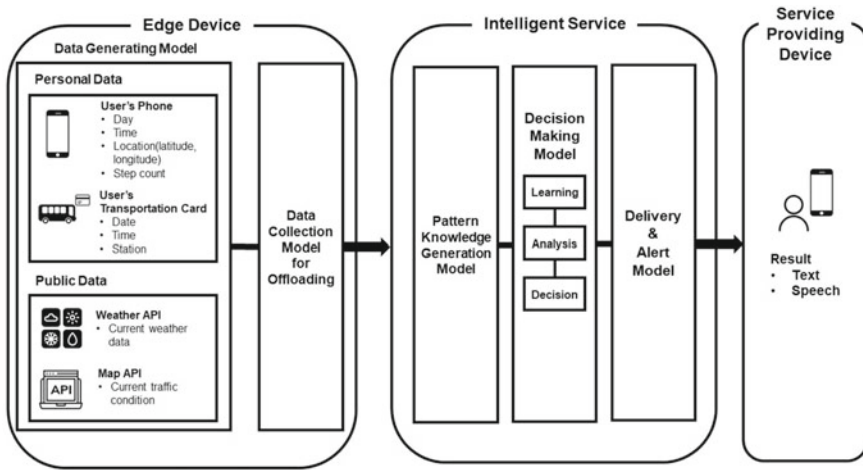


Fig. 2 Architecture of intelligent personalized transport alert system

on weather and traffic conditions. The system provides real-time arrival information and transportation information that match with the user’s daily routine through speech and text-based push message.

Intelligent Personalized Transport Alert System is composed of Edge Device, Intelligent Service, and Service Providing Device as shown in Fig. 2. The data which is generated and offloaded from Edge Device, goes through Intelligent Service and corresponding result is provided to the user in the form of speech and text-based notification by Service Providing Device.

2.1 Edge Device

Edge Device contains two models: Data Generating Model and Data Collection Model. Data Generating Model generates data, and then this data goes through offloading process in Data Collection Model. After the offloading process, data is preprocessed and it enters the server, preventing the server from overloading and therefore reducing the execution time. Sections 2.1.1 and 2.1.2 explains the details of Data generating Model and Data Collection Model.

2.1.1 Data Generating Model

There are two types of generated data. First, personal data is generated from the user’s mobile device and transportation card. Second, public data is provided by API. Public data that was used is weather data and traffic data. The generated data goes through the offloading process.

2.1.2 Data Collection Model for Offloading

In Data Collection Model, data is processed and gets ready to be patterned. Offloading is an effective way to reduce the burden on the server. Offloading preprocesses the data in the Edge Device. By preprocessing the data in the Edge Device, the server gets refined data and this reduces the execution time. Offloaded data is delivered to the Pattern Knowledge Generation Model and gets piled up in the database. These data go through the Decision Making Model and is classified into several patterns. Figure 3 displays the data that is accumulated in the database. There are firebase real time data, android data and card data.

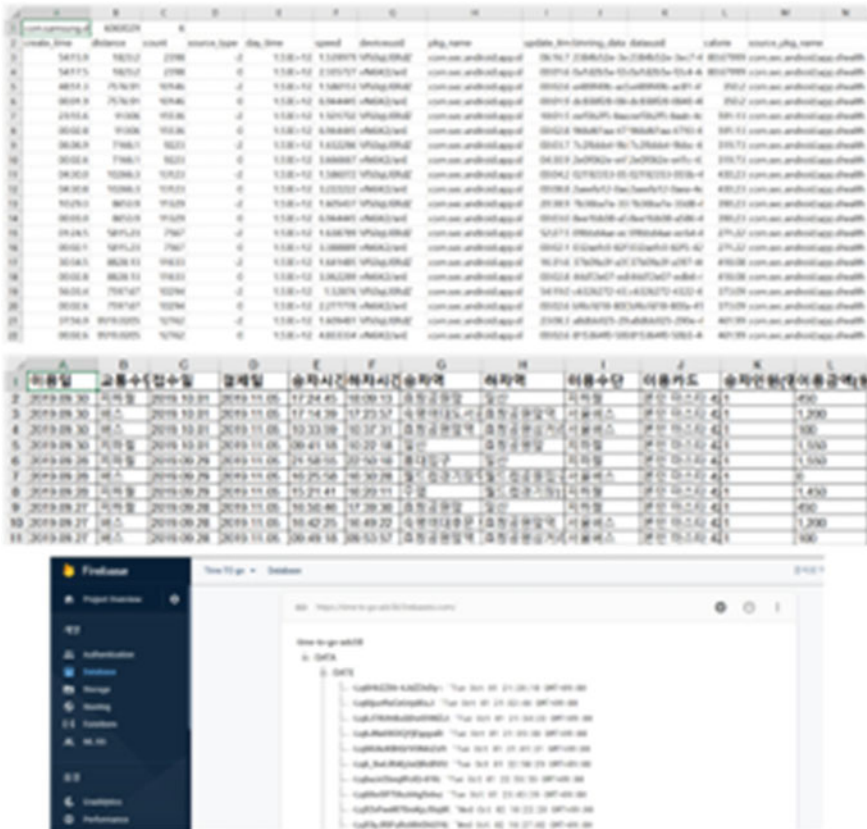


Fig. 3 Accumulated data in the database

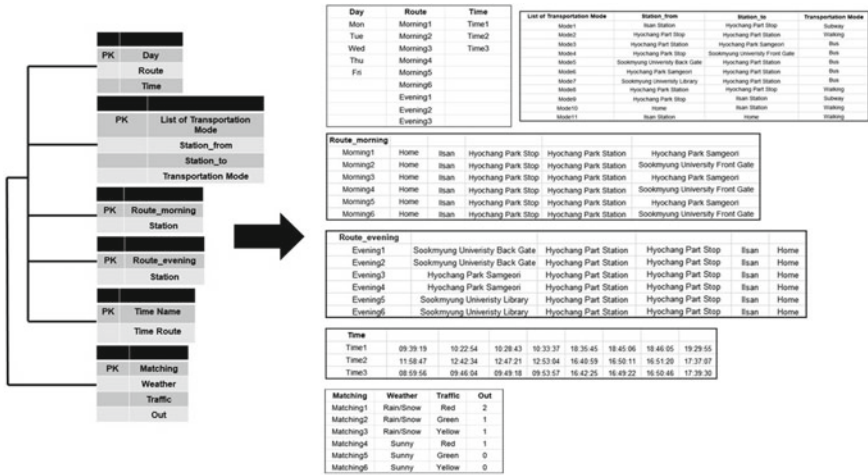


Fig. 4 Database schema and pattern knowledge data

2.2 Intelligent Service

Intelligent Service contains Pattern Knowledge Generation Model, Decision Making Model, and Delivery and Alert Model. Intelligent Service makes a decision based on the data from the Edge Device. First, the Pattern Knowledge Generation Model classifies data into specific patterns. Then, in the Decision Making Model, the pattern goes through a learning process to make decisions. Lastly, the decision is sent to the Delivery and Alert Model which will display the result through speech and text-based notification to the users.

2.2.1 Pattern Knowledge Generation Model

Data stored in the database is processed and is classified into data with specific patterns. The semantics of the pattern is then analyzed and data is stored in the Pattern Knowledge Database according to the pattern. Figure 4 shows the schema of the database and the pattern of data that is processed. Since the data is repeated every week, we classified the data according to the day of the week and the time of the day.

2.2.2 Decision Making Model

Based on the data in the Pattern Knowledge Database, data is trained to make the correct decision. Two models are used in this study. One is for the weekly routine decision, and the other is for weather and traffic situation decision. Data in Pattern

Knowledge Database is used as the training data, and the real time data from android is used as the test data. The result of this model is the list of the most ideal transportation based on the weather and the traffic situation, and the estimated time of arrival.

2.2.3 Delivery and Alert Model

Decision Making Model delivers the decision to the Delivery and Alert Model. Delivery and Alert Model is responsible for delivering the result to the Service Providing Device, and providing notifications to the user's mobile device.

3 Implementation and Analysis

For implementation, Intelligent Personalized Transport Alert System uses the following: Google Firebase, AWS, MySQL, Tensorflow, Linux, and Android Studio.

3.1 Data Gathering

Information on date, time, means of transportation, and corresponding station is gathered using transportation card report. Step count, available on the user's android device, is not saved in database but is sent to the mobile device in real time instead. The arrival time of bus/subway and information on weather and traffic are gathered in real time using Open API provided by public data portal and Kakao API.

3.2 Preprocessing

Preprocessing is performed to process raw data into a suitable form of data. Weather data is preprocessed into 2 types of pattern: Rain/Snow and Sunny. Traffic situation is classified into red, yellow, and green depending on the severity of traffic congestion.

3.3 Machine Learning

Data in the Pattern Knowledge Database is used as the training data, and the real time data from android is used as the test data.

The system has to be trained so that it can notify the users about the transportation information at a certain day and time automatically according to the user's routine that is stored in the Pattern Knowledge Database.



Fig. 5 Result screen

The decision on the most ideal transportation is made based on the analysis of current weather and traffic situation. The combination of the weather data and the traffic situation data determines whether to recommend a bus/subway that the user takes in usual or a bus/subway that arrives faster than usual, to ensure that the user gets to his/her destination on time.

3.4 Development of Mobile Application

The user’s current location and date are received from the user’s mobile device in real time. Step count is also taken in real time from the user’s mobile device, and it is compared with the average step count. The application is developed in a way that it can run in background, since notification is given based on real time location and time. The arrival time of bus/subway depending on the user’s route, is provided in the form of speech and text notification. Speech notification is implemented using Google TTS API. Figure 5 shows the result screen of the application.

4 Conclusion

Intelligent Personalized Transport Alert System can be used in a variety of areas. This system is a mobile application, developed to act as a personal transport guide. It can be used by blind or elderly people who struggle with poor eyesight, due to its ability to provide speech notification. This research mainly focuses on transportation, but when user schedule and health data are provided in addition, it can further be developed into a personal assistant system. Moreover, with more detailed information

on diverse traffic situation and weather, this service could include a function on finding the shortest path to user's destination.

Acknowledgements This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-01456, AutoMaTa: Autonomous Management framework based on artificial intelligent Technology for adaptive and disposable IoT)

References

1. Jung JaeGon (2019) Do it! Android programming. Easyspublishing, Seoul, Republic of Korea
2. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Stanford, California
3. James G, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning: with applications in R. Springer, Stanford, California
4. Sheng QZ, Yu J, Dustdar S (2017) Enabling context-aware web services: methods, architectures, and technologies. CRC Press Florida, Boca Raton
5. Lim BY, Dey AK (2011) Evaluating intelligibility usage and usefulness in a context-aware application, human-computer interaction. *Towards Intell Implicit Interact Part 5*:92–101

Induction Motor Bearing Fault Diagnosis Using Statistical Time Domain Features and Hypertuning of Classifiers



Rafia Nishat Toma and Jong-Myon Kim

Abstract Condition monitoring of induction motors plays a significant role in avoiding unexpected breakdowns and reducing excessive maintenance costs. In the majority of cases, bearing faults are found to be an issue in the failure of induction motors. The detection and valuation of irregularities at an early stage can help prevent disastrous failures. In this paper, the detection and classification of bearing faults in an induction motor are performed using machine learning techniques. The current signal from two different phases is recorded for three motor conditions: healthy, inner race fault and outer race fault. The statistical features are then applied for dimensionality reduction. Finally, the statistical features are used as the input of classifiers, including support vector machines (SVMs), random forests (RFs), and k-nearest neighbor (KNN). The grid search method is used to estimate the best-suited meta-parameters for each classifier to achieve the best performance in fault classification. With the regularization parameters, all the classifiers achieve over 98% classification accuracy.

Keywords Induction motor · Bearing fault diagnosis · Statistical features · Classifiers · Grid search method

1 Introduction

In industrial applications, induction motors are extensively used because of their high reliability and low maintenance characteristics. The main components of an induction motor are the rotor, the stator, the magnets in the permanent magnet synchronous machine (PMSMs), the bearings, and the shaft. These machines work under conditions with severe mechanical, electrical, and thermal stress. Components can be

R. N. Toma

Department of Electrical and Computer Engineering, University of Ulsan, Ulsan, Korea

e-mail: rafiatoma.eceku@gmail.com

J.-M. Kim (✉)

Department of IT Convergence, University of Ulsan, Ulsan, Korea

e-mail: jongmyon.kim@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_35

damaged due to overloading, abrasion, or unbalanced load [1]. Fault diagnosis of induction motors is crucial to avoid losses. Continuous condition monitoring can ensure uninterrupted operation. To address this issue, data-driven condition monitoring is gaining popularity. In this approach, machine learning algorithms are used to build models using historical data. This data is collected from various sensors installed in the motor.

One of the approaches most often used is a signature extraction-based approach in which fault signatures are analyzed in the time domain, the frequency domain, or the time–frequency domain [2]. In the analysis of the time domain, root mean square (RMS) values, peak-to-peak amplitude, and higher order statistics are used [3], while envelop analysis and higher-order spectral analysis are applied for frequency domain analysis [4]. Time–frequency domain analysis contains wavelet transform, Fourier transform and Hilbert transform methods [5]. The extracted signal, including current, voltage, vibration, power, temperature, and acoustic emission is considered as a monitoring signal. Other parameters, such as active and reactive power, thermal field, and magnetic flux [6] are used for fault diagnosis of squirrel-cage induction motors (SCIMs). Recently, the most popular method for machine health monitoring is the vibration signal, because of its facility to convey inherent information of the mechanical system [7]. This process requires external sensors, which are costly, difficult to set up appropriately, and do not guarantee continuous online monitoring. In contrast, the motor current signal analysis (MCSA) technique offers several advantages. It is a spectral analysis method for fault analysis of an induction motor that provides the ability to perform remote monitoring. Furthermore, it does not require any additional sensors. Thus, it is cost effective. Another type of approach is known as the model-based approach, in which a mathematical model is used for predicting the fault condition of an induction motor. There is also an approach known as a knowledge-based approach, which does not need any mathematical model or trigger threshold to identify faults. In this approach, machine learning algorithms and artificial intelligence methods are applied for fault diagnosis for various nonlinear and complex time-varying systems [2]. Among the various machine learning methods, using an artificial neural network (ANN) merged with other techniques is reported in [8].

Among various parts of a motor, we considered a bearing-related fault for this study. There exist two types of bearings, depending on their installation position relative to the motor; there are internal and external bearings. In this paper, machine learning algorithms such as SVM, RF, and KNN are investigated to propose the best model for bearing fault diagnosis of an induction motor. Statistical features, which are used for this model, are derived from motor current signals of two different phases. In addition, performance measures, including precision, recall, confusion matrix, and F-measure are utilized for model evaluation.

This paper is organized as follows. In Sect. 2, characteristics of motor bearing fault diagnosis with current signal analysis are summarized. In Sect. 3, the methodologies for fault feature combination and fault classification are explained briefly. The results for the presented method and the conclusion are presented in Sects. 4 and 5, respectively.

2 Motor Bearing Fault Diagnosis with Current Signature Analysis

Among all fault types in induction motors, faults related to bearing failures occur in approximately 50% of the cases [9]. The key components of induction motor ball bearings are the ball, the outer raceway, and the inner raceway, along with the requirement that there must be uniform distance between the balls to avoid contact with each other, typically accomplished with the help of a cage. The flow of current due to a bearing defect in a motor can be expressed as follows:

$$i_F(t) = i_H(t)[1 + \beta \cos(\omega_c t)] \tag{1}$$

where β , $i_H(t)$ and $i_F(t)$ denote the modulation index, and the motor phase current in healthy and faulty bearing conditions, respectively.

Equation (1) for bearing faults can be represented with consideration of the higher harmonics as:

$$i_F(t) = i_H(t)[1 + \sum_{n=1}^{\infty} \beta \cos(n\omega_c t)] \tag{2}$$

The current of an ideal induction motor for a healthy bearing condition can be represented as [6, 7]:

$$i_H(t) = I_m \cos(2\lambda f_s t) = I_m \cos(\omega_s t) \tag{3}$$

where f_s is the supply frequency.

From Eqs. (2) and (3), the current of an ideal IM with a bearing defect can be written as:

$$i_F(t) = I_m \cos(\omega_s t) + \frac{I_m}{2} \sum_{n=1}^{\infty} \beta_n [\cos((\omega_s - n\omega_c)t) + \cos((\omega_s + n\omega_c)t)]. \tag{4}$$

From (4), it can be assumed that the bearing defect is found at certain frequencies based on its stator current equation, given as:

$$F_{\text{bng}} = f_s \pm n f_c \quad n = 1, 2, \dots \tag{5}$$

3 Methodology

The dataset we use in this work was collected from Kat Data Center of the Chair of Design and Drive Technology, Paderborn University, Germany [9]. In this work, two current signals with a phase difference of 180° are considered. The current signals are taken for 17 different combinations of torque, speed, and angular velocity under three bearing conditions: healthy bearings, damage in the inner ring, and damage in the outer ring. Initially, the data from every combination is concatenated and the classes are labeled as 0, 1, and 2 for healthy bearings, inner ring faults, and outer ring faults, respectively. The workflow is given in Fig. 1.

In the next stage, 10 different statistical features (mean, median, standard deviation, variance, sum, skewness, kurtosis, energy, RMS, and crest factor) are extracted from the dataset. Among them, the mean and variance are employed to define the probability density function of the time-varying signal. The skewness parameter helps to measure the distribution symmetry. Other higher-order statistical features are also analyzed, such as kurtosis, which helps in comparing the divergence between mean by a minute value with those with a high value of divergence.

In this work, three widely used and popular machine learning algorithms, support vector machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN), are used to build a classification model. Each algorithm has different hyperparameters. Selecting the best value for a hyperparameter is important for maximizing performance.

For an SVM, the cost (C), kernel type, and kernel width parameter (γ), or Gamma, are the three important parameters. The cost parameter determines the extent of misclassification, which can be allowed for non-separable training data. Sometimes a higher value of C can create an overfitting problem. Therefore, to avoid overfitting (as well as underfitting), selecting an optimal value for C is critical. On the other hand, the shape of the hyperplane is determined by the value of Gamma. If a high value of gamma is chosen, the SVM tries to separate every train data point, and therefore the decision boundary becomes very curvy. In such a case, the model might not perform well for test data points. RF is an ensemble learning algorithm that works by constructing a multitude of decision trees during the training period. The number of trees, the maximum depth, the number of features in every split, and the number of sample leaves are the important hyperparameters of RF. Finally, for the k-NN algorithm, the critical hyperparameter is the number of neighboring samples it

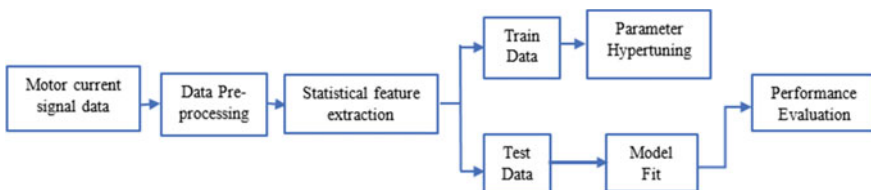


Fig. 1 Workflow diagram

considers. For selecting the appropriate value of the hyperparameters the GridSearch method is used in this work. The dataset is divided into two different ratios, such as 70:30 and 80:20, for training and testing. There are 17 subsets for various load conditions of induction motors, and the train-test ratio is maintained for every load condition.

4 Results

There exists many metrics to assess the performance of a classification algorithm. In this work, the overall accuracy, confusion matrix, prediction, recall, and F1 for 3 different classes are considered as performance metrics. The precision, recall, F-1 score, and overall accuracy are derived from the confusion matrix, as shown in the following equations:

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive} + \text{False Positive (FP)}} \tag{6}$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive} + \text{False Negative (FN)}} \tag{7}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{9}$$

Precision is an effective measure for the case in which the cost of FP is high. On the other hand, recall calculates the actual TP values and becomes the selection metric for the best model in the case of a high cost with FN. However, the F1 score presents a harmonic mean of the precision and recall, taking both metrics into account. The confusion matrices for SVM, RF, and KNN are provided in Fig. 2.

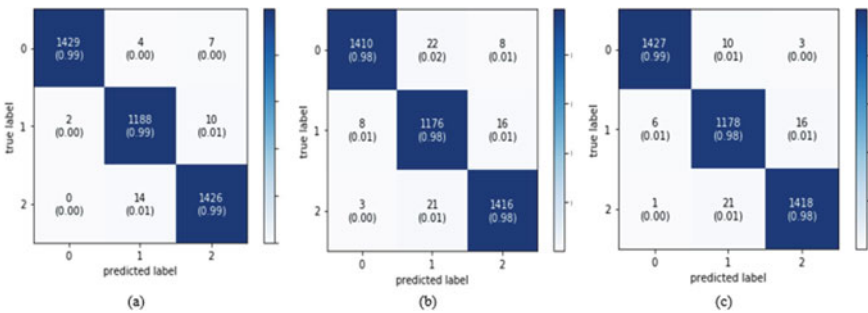


Fig. 2 Confusion matrix for a SVM, b RF, and c KNN

Table 1 Hypertuning parameters

ML algorithms	Hyper tuned values
SVM	C = 10, gamma = 1, kernel = rbf
RF	Random state: 42 Max features: sqrt Number of estimators: 100 Maximum depth: 10
KNN	Leaf size: 1 Number of neighbors: 10 Weights: uniform

Table 2 Result for 3 different ML classifier

	Precision	Recall	F1	Accuracy
SVM	0.99	0.99	0.99	0.99
RF	0.97	0.99	0.98	0.98
KNN	0.99	0.98	0.99	0.99

Table 1 represents the optimum value of the hyperparameters of the three learning algorithms, obtained from the GridSearch method.

The value of the performance parameters are shown in Table 2. The overall accuracy rate for all three algorithms is very high. SVM and RF demonstrated similar performance levels. Each model has high precision and accuracy values, which indicates that the model is classifying each class accurately.

5 Conclusion

In this paper, a novel fault detection and classification method for induction motors using current signal data was presented. To obtain the current signal data, the MCSA method was utilized. Ten statistical features for two different phase currents were evaluated for the healthy condition and two faulty conditions, and fed as an input to the SVM, RF, and KNN classifiers. We observed that all three classifiers using the statistical features as input performed well on the current signal data. All the performance measures, including precision, recall, F1 score, and accuracy showed high performance for the three classifiers. In the future, we will focus on the implementation of frequency domain analysis along with machine learning algorithms for detecting and classifying faults in induction motors.

Acknowledgements This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20181510102160, No. 20192510102510).

References

1. Mbo'o CP, Hameyer K (2016) Fault diagnosis of bearing damage by means of the linear discriminant analysis of stator current features from the frequency selection. *IEEE Trans Ind Appl* 52(5):3861–3868
2. Ali MZ, Shabbir MNSK, Liang X, Zhang Y, Hu T (2019) Machine learning-based fault diagnosis for single- and multi-faults in induction motors using measured stator currents and vibration signals. *IEEE Trans Ind Appl* 55(3):2378–2391
3. Siegel D, Ly C, Lee J (2012) Methodology and framework for predicting rolling element helicopter bearing failure. *IEEE Trans Reliab* 61(4):846–857
4. Randall RB, Antoni J (2011) Rolling element bearing diagnostics—a tutorial. *Mech Syst Signal Process* 25(2):485–520
5. Cabal-Yepez E, Garcia-Ramirez AG, Romero-Troncoso RJ, Garcia-Perez A, Osornio-Rios RA (2013) Reconfigurable monitoring system for time-frequency analysis on industrial equipment through STFT and DWT. *IEEE Trans Industr Inf* 9(2):760–771
6. Cabanas MF et al (2011) A new portable, self-powered, and wireless instrument for the early detection of broken rotor bars in induction motors. *IEEE Trans Industr Electron* 58(10):4917–4930
7. Jing L, Zhao M, Li P, Xu X (2017) A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement* 111:1–10
8. Boukra T, Lebaroud A, Clerc G (2013) Statistical and neural-network approaches for the classification of induction machine faults using the ambiguity plane representation. *IEEE Trans Industr Electron* 60(9):4034–4042
9. Lessmeier C, Kimotho JK, Zimmer D, Sextro W (2016) Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark data set for data-driven classification. In: *Proceedings of the European conference of the prognostics and health management society*, vol 7, pp 05–08

Crack Detection Using Fully Convolutional Network in Wall-Climbing Robot



Myeongsuk Pak and Sanghoon Kim

Abstract Since the wall crack inspection of structures is difficult to access and to secure objectivity of the visual inspection, research on automatic inspection is being conducted. This paper is a study on the automatic detection of wall cracks of the wall-climbing robot, which aims to detect the cracks by robot itself in real time. Deep learning techniques are also applied to crack inspection, but there are difficulties in resource limitation in embedded environments. In this study, we examined the performance by experimenting with deep learning method that can be applied to embedded environment and the possibility of applying them to wall-climbing robot was presented.

Keywords Crack detection · FCN · Wall-climbing robot

1 Introduction

Crack inspection of large structures, such as high-rise buildings and bridges, is dangerous for human access and is time-consuming and expensive, including the installation of additional structures. In addition, it is difficult to secure the objectivity of visual examination. In order to solve this risk and efficiency problem, many researches have been done on the automatic crack detection technique using a robot. In the case of the crack detection technique using UAV (drone) [1, 2], which is the most active research, there is a disadvantage in that it cannot be photographed close to the wall due to the danger of the UAV itself. Some studies using ground robots [3], however, have the disadvantage that they cannot be driven on walls of structures such as buildings and bridges. In this paper, we focus on wall crack detection using deep learning in wall-climbing robot equipped with high performance embedded platform and webcam. In the case of the wall mobile robot, it is attached to the wall

M. Pak · S. Kim (✉)

Department of Electrical, Electronic and Control, Hankyong National University, 327, Jungang-ro, Anseong, Gyeonggi-do, Korea
e-mail: kimsh@hknu.ac.kr

and moves, so it is possible to detect the micro cracks and to drive the indoor and the bridge.

The distribution and shape of cracks on the surface of structures is irregular and linear, which can be seen as edge or line objects. For greater accuracy, crack detection is treated as semantic segmentation that classifies at pixel level to obtain crack location. In high quality images with good continuity and high contrast, cracks can be detected with high accuracy through traditional methods such as Sobel and Canny edge detection [4]. However, due to the high influence of lighting and wall material, cracks may be less continuity due to noise and lower contrast.

Recently, researches to improve the accuracy of crack detection using deep learning techniques have been conducted. Liu et al. [5] proposed a method for concrete crack detection that achieves high accuracy with a smaller training set based on U-Net that is more robust, effective and more accurate. Zhang et al. [6], inspired by Full Convolutional Networks (FCN), proposed a full convolutional network based on dilated convolution consisting of encoders and decoders, demonstrating faster convergence and better generalization in concrete crack test sets. Zou et al. [7] proposed DeepCrack for crack detection based on SegNet architecture. They effectively infer the crack by pairwise fusing the convolutional features generated in the encoder and decoder networks at the same scale. These methods were experimented with GPU on desktop computer and are very accurate using VGG16, Resnet18, etc. as based network, but the speed is difficult to apply to the mobile environment.

In order to apply deep learning to mobile robots, a tradeoff between performance and speed is required. In this study, the performance is tested and the results are presented by applying the deep learning method that can perform crack detection in real time on mobile robots with limited resources that can perform well on non-homogeneous walls such as uneven concrete surfaces.

2 Crack Detection Method

The network architecture used for semantic segmentation consists of two main components: an encoder responsible for feature extraction and a decoder that calculates the final class probability through upsampling. In this section, we look at the state-of-the-art semantic segmentation techniques and discuss the underlying algorithms suitable for mobile computing environments to drive deep learning crack detection in wall-climbing robots. Segmentation networks such as SegNet and DeepLab have high computational costs and requirements for specific hardware. Since our main limitation is speed, ShuffleSeg and DeepLabv3+ are chosen.

2.1 *ShuffleSeg*

ShuffleSeg [8] is a computationally efficient segmentation network that reduces computational cost while maintaining good accuracy based on grouped convolution and channel shuffling in the encoder. It delivers 58.3% mIoU in the CityScapes test set and runs at 15.7 frames per second on NVIDIA Jetson TX2. The encoder is based on ShuffleNet and uses skip connection FCN8s to the decoder. For training, weighted cross entropy loss is used, the weight decay is $5e-4$ and the Adam optimizer is used with learning rate $1e-4$.

2.2 *DeepLabv3+*

DeepLabv3+ [9] combines the advantages of the spatial pyramid pooling module and the encoding decoder architecture, and extends DeepLabv3 by adding a simple but effective decoder module. Apply depthwise separable convolution to the Atrous Spatial Pyramid Pooling and Decoder module to create a faster, more powerful encoder-decoder network. Using the Xception model, the PASCAL VOC 2012 and Cityscapes datasets achieved mIoU of 88.9% and 82.1%, respectively. In their implementation we use MobileNet to fit the mobile platform.

3 Experiments

3.1 *Experimental Environment*

NVIDIA Jetson TX2 is installed to apply deep learning to the wall-climbing robot shown in Fig. 1. In this experiment, the model was tested using Jetson TX2, whose specifications are shown in Table 1.

3.2 *Crack Data Training*

The crack image used in this study is based on the data set of [10] and contains about 11,200 images that are merged from 12 crack segmentation datasets. The crack image consists of 9,603 training images and 1,695 test images, with a size of 448×448 . For the semantic segmentation, an annotation image was used by changing the background to 0 and the crack to 1. Training of the crack data was performed on a desktop computer with an Intel Core i7-8700 K CPU 3.7 GHz CPU and NVIDIA Geforce 2080.

Fig. 1 Prototype of our wall-climbing robot



Table 1 Jetson TX2 specification

CPU	GPU	Memory
Dual-core Denver 2 64-bit CPU and quad-core ARM A57 complex	NVIDIA Pascal™ architecture with 256 NVIDIA CUDA cores	8 GB

3.3 Results

Two segmentation networks were tested for crack detection. Figure 2 shows the crack detection results. In the case of DeepLabv3+ , only part of the thin crack is detected, as in the first column, and some crack-like patterns are detected, as in the second column. For ShuffleSeg, both thin and coarse crack were detected well. ShuffleSeg has a speed of 18 fps with a test images on the Jetson TX2, and is suitable for wall-climbing robots.

4 Conclusion

In this paper, we use semantic segmentation technique to classify cracks in pixel level for crack detection of wall-climbing robot. In the case of the wall-climbing robot, it is attached to the wall and moves, so it is possible to detect the micro cracks and to drive the indoor and the bridge.

We tested the performance by applying a deep learning technique that can perform good performance even on non-homogeneous walls such as non-smooth concrete surfaces and detect cracks in real time on mobile robots with limited resources. We showed that we can run the crack detection algorithm using deep learning inside the robot. Further work will require more experiments to improve detection accuracy

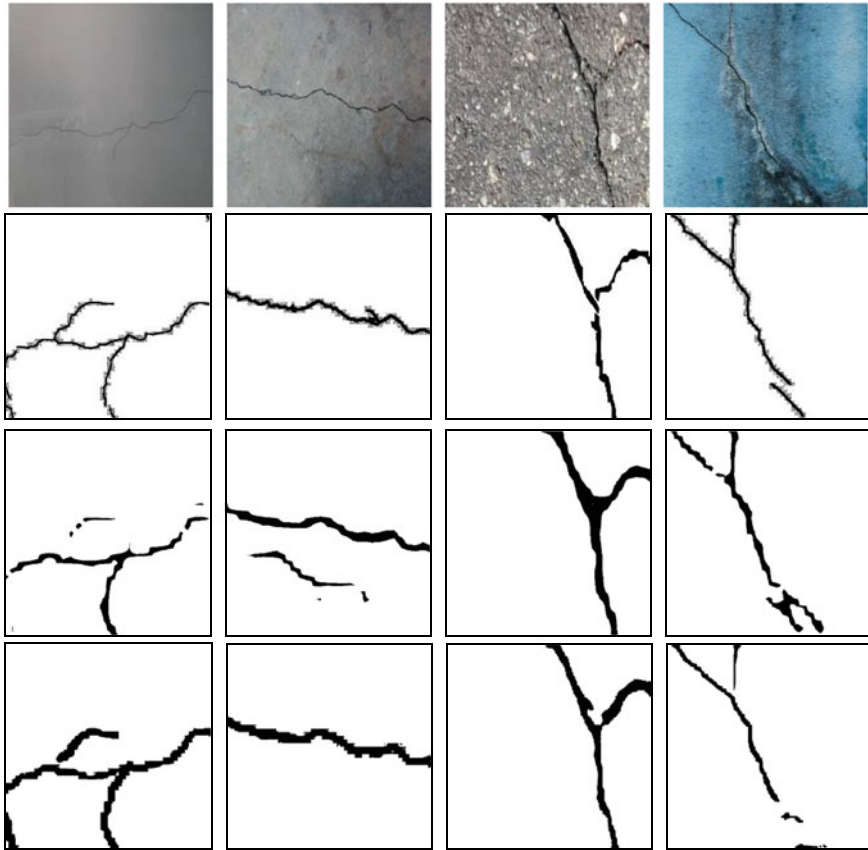


Fig. 2 Crack detection results. The first row is the input image, the second row is the ground truth, the third row is the result of DeepLabv3+ and the last row is the result of ShuffleSeg

and better detection efficiency, and it is necessary to improve the existing methods to increase crack detection performance on wall-climbing robots.

References

1. Kim JW, Kim SB, Park JC, Nam JW (2015) Development of crack detection system with unmanned aerial vehicles and digital image processing. In: World congress on advances in structural engineering and mechanics
2. Kim H, Lee J, Ahn E, Cho S, Shin M, Sim SH (2017) Concrete crack identification using a UAV incorporating hybrid image processing. *Sensors* 17(9):E2052
3. An S, Jang J, Han C, Kim P (2004) An inspection system for detection of cracks on the concrete structures using a mobile robot. In: 21st international symposium on automation and robotics in construction, Korea, pp 21–25

4. Abdel-Qader I, Abudayyeh O, Kelly ME (2003) Analysis of edge-detection techniques for crack identification in bridges. *J Comput Civil Eng* 17(4):255–263
5. Liu Z, Cao Y, Wang Y, Wang W (2019) Computer vision-based concrete crack detection using U-net fully convolutional networks. *Autom Construct* 104:129–139
6. Zhang J, Lu C, Wang J, Wang L, Yue XG (2019) Concrete cracks detection based on FCN with dilated convolution. *Appl Sci* 9(13):2686
7. Zou Q, Zhang Z, Li Q, Qi X, Wang Q, Wang S (2018) Deepcrack: learning hierarchical convolutional features for crack detection. *IEEE Trans Image Process*, pp 1–15
8. Gamal M, Siam M, Abdel-Razek M (2018) ShuffleSeg: real-time semantic segmentation network. [arXiv:1803.03816](https://arxiv.org/abs/1803.03816)
9. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*
10. Concrete Crack Images for Classification, https://github.com/khanhha/crack_segmentation#Dataset

Performance Evaluation of AODV and AOMDV Routing Protocols Under Collaborative Blackhole and Wormhole Attacks



Tran Hoang Hai, Nguyen Dang Toi, and Eui-Nam Huh

Abstract Mobile Ad hoc Network (MANET) is easy to be attacked than wired networks due to its characteristics of open network topology, high mobility, lack of physical security and independent management. This paper analyzes the impact of routing attacks on the performance of the AODV and AOMDV routing protocol under several collaborative routing attack scenarios (blackhole and wormhole). We conclude that AOMDV is better than AODV in case of collaborative attacks and we show that collaborative Blackhole and Wormhole attacks can affect much performance of the network, especially in the case when we do not recognize the location area of malicious nodes.

Keywords MANET · Routing attack · Network security · Blackhole · Wormhole

1 Introduction

Mobile Ad Hoc Networks (MANET) is made up of dynamically, self-configuration nodes in a flat network without any fixed or existing infrastructure and centralized administrator. In MANET, the network nodes share a wireless channel without any centralized control or management. Each node in MANET works as both end host and router at the same time. There are several routing protocols designed in MANET such as AODV, DSR, OLSR, etc. which can be classified by proactive routing protocols, reactive routing protocols, and hybrid routing protocols which is the combination of

T. H. Hai (✉) · N. D. Toi

School of Information and Communication Technology, Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam
e-mail: hai.tranhoang@hust.edu.vn

N. D. Toi

e-mail: 20153860@student.hust.edu.vn

E.-N. Huh

Department of Computer Science and Engineering, Kyung Hee University, Deogyong-daero, Giheung-gu, Yongin-si, Gyeonggi-do, South Korea
e-mail: johnhuh@khu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_37

273

the previous two [1–3]. In several works have been studied in the last few years [4–6], the proactive routing protocols provide better performance (in terms of quantitative metrics such as throughput, packet delivery ratio, end to end delay, etc.) than reactive protocols. In this paper, we focus on proactive protocols since the hybrid routing protocols usually use much computation resources rather than both proactive and reactive protocols [7]. Several attacks in network layer have been identified and studied in security research of MANET in [8], such as blackholes, wormholes, link spoofing, gray holes, link spoofing, etc. There are related works of how an attack can affect the operation of MANET, but we focus on more complicated attack scenarios, i.e. mobility attacker, several attacks deploying simultaneously, etc. Understanding how MANET works and analyzing its routing performance under attacks are always the first important tasks to project the overall system.

2 Related Works

In MANET, an attacker can re-route network traffic, or inject itself into the path between the source and destination and thus control the network traffic flow [8]. Several attacks have been identified and studied in security research [9–15]. One of the most common routing attacks in MANET is the Blackhole attack. In this attack, a “black” node within the network displays itself as having the shortest path to the destination node. Once the packets are drawn to the attacker, they are then dropped instead of relayed, and the communication of the MANET will be disrupted. In wormhole attacks, the attacker receives packets at one point in the network and tunnels them to another part of the network for malicious purposes. In MANET with AODV routing protocol, this attack can be done by tunneling every REQUEST to the target destination node directly. When the destination’s neighboring nodes hear this REQUEST packet, they will rebroadcast that REQUEST packet in a normal operation and then discard any other REQUESTS for the same route discovery [16]. There is a huge work on the study of how blackhole and wormhole can manipulate the network traffic in MANET but mostly the authors focus on separate, single and static routing attack. In [11], the authors analyzed the performance of AODV and OLSR protocols individually with only one black hole attack and without blackhole attack. In [17], the authors analyzed the performance of Mobile Ad hoc Networks (MANET) under black hole and wormhole attack separately for AODV protocol. In [18], AODV and DSDV protocols are analyzed in terms of routing overhead, packet delivery ratio, throughput and end to end delay under single Black hole attack and collaborative Black hole attacks.

3 Collaborative Blackhole and Wormhole Attacks

Over the years, many researchers have analyzed the performance of MANET networks against attacks of specific types of attacks. However, in reality, an active MANET network can face many types of attacks at the same time. Those types of attacks not only affect normal nodes but also directly impact others. In this paper, we propose analyzing the performance of a MANET network that runs AODV and AOMDV protocols in case of collaborative blackhole and wormhole attacks. An attacker is only concerned with packet loss and performance degradation of the MANET network, so Blackhole will be preferred due to its simple operation mechanism, less energy and computation costs than wormholes. How a Blackhole attack can affect the network also depends on operation area of malicious nodes, their positions compared to others. When attackers want to target a large-scale attack on the network at a given time, i.e. collaborative of Blackhole and wormhole, it is necessary to distribute malicious nodes into MANET so that they can interfere with the entire routing process. However, the unique feature of the blackhole is that it always drops packets until the nodes in the network detect the anomalies which reduces the ability to hide themselves. A smart strategy for an attacker is to allocate time for collaborative attacks to hide malicious nodes in the network, combining both types of attacks will make the attack strategy more flexible by introducing Wormhole.

4 Simulation Results

4.1 Simulation Scenarios

In this paper, we focus on analyzing the performance of MANET running AODV and AOMDV routing protocols under collaborative Black hole and wormhole attacks. The simulation uses ns2.35 running on Ubuntu 14.04. The network environment is a 1200×800 plane with the number of nodes is 50, 80, 100, and 120 respectively. Since assuming that we focus on the performance of routing protocols, the semantic scope in the wormhole attack can be ignored and nodes considered wormhole link have a *drop factor* K , in this case $K = 40$. For each routing protocol, to know details of the operation under collaborative attacks, we focus on three main scenarios follows:

- Script 1: The normal nodes and malicious nodes are evenly distributed and fixed in the analysis plane.
- Script 2: The nodes are evenly distributed in the analysis plane but the malicious nodes move evenly diagonally from the analysis plane.
- Script 3: The normal nodes and malicious nodes are randomly placed in the analysis plane.

For more details, each scenario will be simulated in three subscripts follows:

- Subscript 1: The network only consists of normal nodes.
- Subscript 2: The network consists of normal nodes and two blackhole nodes.
- Subscript 3: The network consists of normal nodes, two blackhole nodes and one wormhole tunnel.

4.2 Simulation Results

In the beginning, first we look on the packet delivery ratio of the 1st scenario when the normal nodes and malicious nodes are evenly distributed and fixed in the analysis plane. We can see in Fig. 1 that when the network only consists of normal nodes, the performance of both routing protocols is very good. When network having blackhole nodes, the packet delivery ratio is decreasing but AOMDV provides better results than AODV in this case. The similar results can be seen with AOMDV when network having two Black hole nodes and one Wormhole tunnel.

We can see in Fig. 2 that in general, AOMDV provides better End-to-end delay than AODV. In the case of network having 80 nodes with AOMDV, the End-to-end delay increases abnormally due to its location characteristics leading to the nodes which cannot find redundant routes leading to fail link of discovery process. In the results of Fig. 3, AOMDV provides better Throughput than AODV generally but the result is not so good under collaborative Blackhole and Wormhole attacks, but it is improving along with network density. Now we look at the results of 2nd simulation scenario in Fig. 4 when nodes are evenly distributed in the analysis plane, but the malicious nodes move evenly diagonally from the analysis plane. We can see that AOMDV still provides better Packet Delivery Ratio than AODV even in

Fig. 1 Packet delivery ratio of 1st simulation scenario



Fig. 2 End-to-end delay of 1st simulation scenario

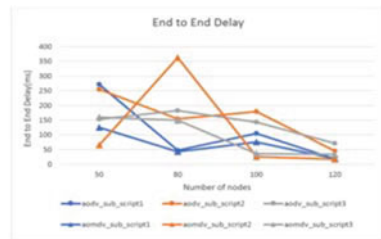


Fig. 3 Throughput of 1st simulation scenario

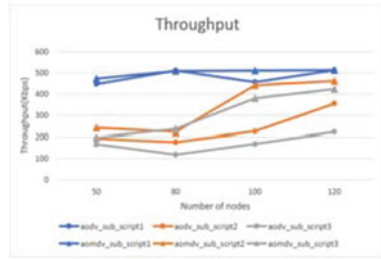
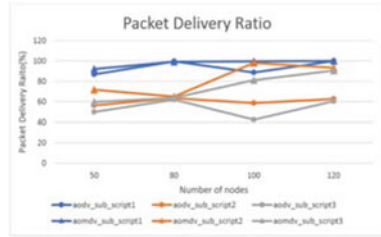


Fig. 4 Packet delivery ratio of 2nd simulation scenario



collaborative routing attacks. Due to the mobility of malicious nodes, the links in the network will be unstable leading to an increase in discovery process. As the result, the increasing in the number of routing packets leading to higher load and computation time. Therefore, End-to-end delay will be higher in a network containing malicious nodes in Fig. 5. In collaborative Blackhole and Wormhole attacks, AOMDV still provides good throughput if the network density is high as in Fig. 6. Now we look

Fig. 5 End-to-end delay of 2nd simulation scenario

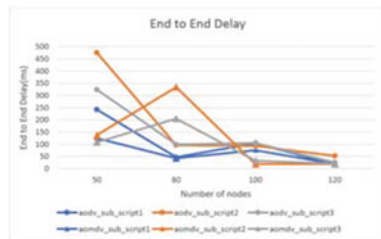
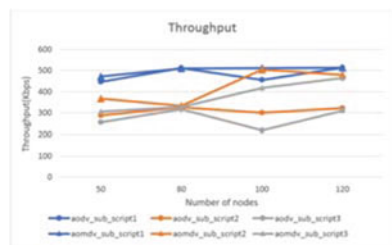


Fig. 6 Throughput of 2nd simulation scenario



at the results of 3rd simulation scenario when normal nodes and malicious nodes are randomly placed in the analysis plane. When all nodes are randomly considered in the 1200×800 plane, the ratio depends on the location of malicious nodes, source node and destination node location and network density in Fig. 7. Because the random algorithm of ns2.35 has not been optimized, the locations of the nodes are not distributed evenly on the plane but the higher density in some areas. It leads to a higher malicious node influence due to the presence of many neighbors. Therefore, the node must process more information due to the high number of incoming packets resulting in high latency in Fig. 8. Therefore, throughput result of this simulation scenario is not promising under collaborative attacks, as in Fig. 9.

Fig. 7 Packet delivery ratio of 3rd simulation scenario

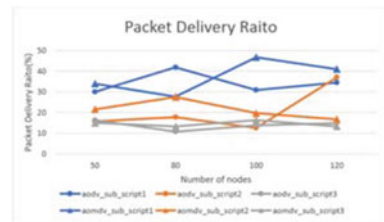


Fig. 8 End-to-end delay of 3rd simulation scenario

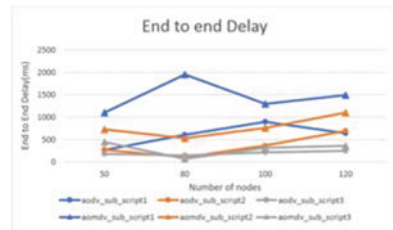
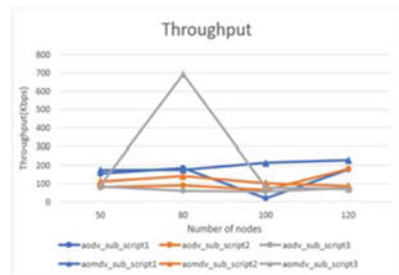


Fig. 9 Throughput of 3rd simulation scenario



5 Conclusion

In this paper, we have shown that collaborative Blackhole and Wormhole attacks can affect performance of the network, especially in the case we do not recognize the random malicious nodes. In general, AOMD provides better results than AODV in most of the cases when the network survives under attacks.

Acknowledgements This work was supported by Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00294, Service mobility support distributed cloud technology).

References

1. Yadav R, Rao S (2015) A Survey of various routing protocols in MANETs. (IJCSIT) Int J Comput Sci Inf Technol 6(5):4587–4592
2. Shruthi S (2017) Proactive routing protocols for a MANET—a review. In: 2017 International conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC), Palladam, pp 821–827
3. Patel DN, Patel SB, Kothadiya HR, Jethwa PD, Jhaveri RH (2014) A survey of reactive routing protocols in MANET. In: International conference on information communication and embedded systems (ICICES2014), Chennai, pp 1–6
4. Istikmal V, Leanna Y, Rahmat B (2013) Comparison of proactive and reactive routing protocol in mobile adhoc network based on “Ant-algorithm”. In: 2013 International conference on computer, control, informatics and its applications (IC3INA), Jakarta, pp 153–158
5. Govindasamy J, Punniakody S (2018) A comparative study of reactive, proactive and hybrid routing protocol in wireless sensor network under wormhole attack. J Electr Syst Inf Technol 5(3):735–744. ISSN 2314-7172
6. Bai Y, Mai Y, Wang N (2017) Performance comparison and evaluation of the proactive and reactive routing protocols for MANETs. In: 2017 Wireless telecommunications symposium (WTS), Chicago, IL, pp 1–5
7. Raheja K, Maakar SK (2014) A. Survey on different hybrid routing protocols of MANET. (IJCSIT) Int J Comput Sci Inf Technol 5(4):5512–5516
8. Joshi P (2011) Security issues in routing protocols in MANETs at network layer. Procedia Comput Sci 3
9. Gurung S, Chauhan (2019) A survey of black-hole attack mitigation techniques in MANET: merits, drawbacks, and suitability. Wirel Netw
10. Fazeldehkordi E, Amiri IS, Akanbi OA (2016) Chapter 2—Literature review. In: Fazeldehkordi E, Amiri IS, Akanbi OA (eds) A study of black hole attack solutions, Syngress, pp 7–57. ISBN 9780128053676
11. Praveen KS, Gururaj HL, Ramesh B (2016) Comparative analysis of black hole attack in ad hoc network using AODV and OLSR protocols. Procedia Comput Sci 85
12. Yaseen QM, Aldwairi M (2018) An enhanced AODV protocol for avoiding black holes in MANET. Procedia Comput Sci 134:371–376
13. Nagrath P, Gupta B (2011) Wormhole attacks in wireless adhoc networks and their counter measurements: a survey. In: 2011 3rd International conference on electronics computer technology, Kanyakumari, pp 245–250
14. Farjammnia G, Gasimov Y, Kazimov (2019) Review of the techniques against the wormhole attacks on wireless sensor networks. Wirel Pers Commun 105:1561

15. Dutta N, Singh MM (2019) Wormhole attack in wireless sensor networks: a critical review. In: Mandal J, Bhattacharyya D, Auluck N (eds) *Advanced computing and communication technologies. Advances in intelligent systems and computing*, vol 702. Springer, Singapore
16. Datta R, Marchang N (2012) Chapter 7—Security for mobile ad hoc networks. In: *Handbook on securing cyber-physical critical infrastructure*. Morgan Kaufmann
17. Kumar M (2013) Analysis of black hole and wormhole attack using AODV protocol. *Int J Res Manag Sci Technol* 1(1):44–28. E-ISSN: 2321-3264. www.ijrmst.org
18. Chavan AA, Kurule DS, Dere PU (2016) Performance analysis of AODV and DSDV routing protocol in MANET and modifications in AODV against black hole attack. *Procedia Comput Sci* 79

Simulation and Analysis of RF Attacks on Wireless SCADA System



Sung-Won Lee, Ji-Hun Kim, and Jonghee Youn

Abstract Those computer systems that are in charge of the control of national and social infrastructures are called Supervisory Control And Data Acquisition (SCADA) systems. Most SCADA systems have been considered to be relatively safe in the past as they use closed networks non-disclosed communication protocols and therefore, the reinforcement of system security and incident responses has been neglected. However, the SCADA system security incidents that have been occurring recently mean that SCADA systems are exposed to security threats and as communication technologies are developed, illegal access paths to SCADA communication networks that use wireless communication can increase further. Since SCADA systems are in charge of the control of national and social infrastructures, the occurrence of security incidents can result in serious problems. Therefore, studies of related security technologies and policies are necessary. To this end, the present paper helps the understanding of the overall system through detailed analyses of the ZigBee protocol among the wireless communication protocols used by the SCADA system. In addition, the security vulnerability of wireless SCADA systems is studied through simulations of products that use the ZigBee protocol.

Keywords SCADA system · ZigBee protocol · Wireless system · RF attack

1 Introduction

The SCADA (Supervisory Control And Data Acquisition) systems [1] that will be mainly dealt with in the present paper are included in major kinds of industrial control systems together with the DCSs (Distributed Control Systems). SCADA systems refer to those systems that control and monitor large scaled distributed

S.-W. Lee · J. Youn (✉)

Department of Computer Engineering, Yeungnam University, Gyeongsan, Republic of Korea

e-mail: youn@yu.ac.kr

J.-H. Kim

Kim & Chang, Seoul, Republic of Korea

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_38

equipment and devices and the distributed control system, that is, the DCS generally controls systems within relatively short distances using communication protocols mutually connected for control [2]. However, after the beginning of the modern times, the differences between the two representative types of industrial control system have become obscure due to the development of information and communication technologies (ICT).

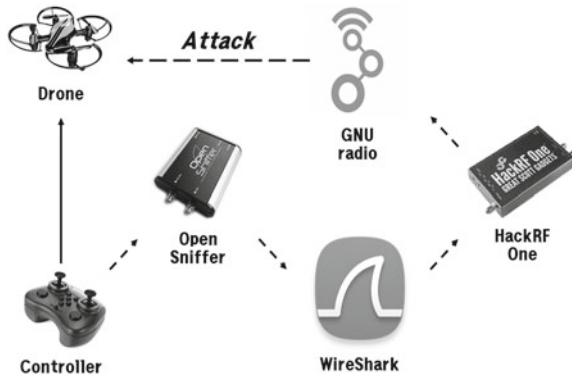
Furthermore, thanks to the recent development of communication technologies SCADA systems applied with wireless communication technologies instead of wire communication have appeared and institutions that operate SCADA system have been introducing increasingly more wireless communication based SCADA systems for the reason of improvement in the efficiency of the access of the control system or control flexibility. However, since most wireless communication based SCADA systems are operated using both wire/wireless communications, they can be said to be more vulnerable to security threats using networks and communication protocols. The purpose of the present paper is to deal with overall security threats against wireless communication based SCADA systems, analyze ZigBee communication among wireless communications, and investigate security threats against SCADA systems through simulations of various equipment units.

2 SCADA System

The architecture of general SCADA systems is largely divided into the Master Terminal Unit (MTU) that plays the role of data collection and control, the Remote Terminal Unit (RTU) in charge of data transmission, the Human-Machine Interface (HMI), and the network [3]. The MTU collects data from the RTU through the communication network and monitors and controls the entire SCADA system's communication based on the collected data. The RTU collects information from data collection devices such as the sensors of the SCADA system in real time and transmits the data to the MTU. The HMI is in charge of the function to visualize and display the collected data in the forms of texts and graphs for efficient communication with the person in charge of operation. The networks of SCADA systems were based on local area network (LAN) or wide area network (WAN) based Ethernet networks in the past. Recently, however, the networks have been configured using diverse communication methods such as Bluetooth, Zigbee, and GPS for wireless communication.

3 Simulation

Although directly studying ZigBee protocol based wireless communication SCADA networks is suitable for the original intent of studies, since most SCADA systems are

Fig. 1 Simulation flow chart

currently used to control major national and social infrastructures because of their nature, there are quite some restrictions on the progression of studies.

In the present paper, instead of security threats to actual SCADA systems, security threats to drones, lighting systems, and smart home kits that use ZigBee communication are demonstrated to identify the attack methods and vulnerabilities of internal networks consisting of ZigBee communication through the demonstration. Figure 1 shows the tool and attack method used in our simulation. The tools used for the demonstration are OpenSniffer [4] and HackRF One [5] and the Wireshark and GNU radio [6] programs are also used for the demonstration. The main flow of the simulation begins with packet sniffing followed by analysis, attack signal generation, and attacks and in addition, simulations of replay attacks and jamming attacks were also carried out.

3.1 Replay Attack [7]

Replay Attacks are an attack issue frequently applied to wireless equipment recently and refers to those attacks that store the communication signals between wireless equipment units as they are and reuse the signals. An advantage of these attacks is that there is no need to analyze signals and these attacks are easy and powerful because the signals once stored when the attack was applied can be continuously used. Exploiting the fact that wireless signals can be sent and received to/from the basic library provided by HackRF One, signals between a drone and a controller can be intercepted to carry out replay attacks and these attacks can control drones because when the stored signals are sent more strongly at shorter periods than those of the controller, the drone receives the attack signals first rather than controller signals. Cases where the channels used when commands are connected between the drone and the controller occur and the changed channels can be identified through the channel scanning provided by the Open Sniffer web interface.

3.2 Jamming Attack [8]

Jamming Attacks refer to those attacks that cause disturbances to wireless signals in the frequency channel area being used so that the equipment cannot carry out the desired operation. Because of the nature of wireless equipment, signals at the same frequency are received first and whether the signals are proper ones or not is judged. Exploiting this fact, the attacker transmits meaningless signals to fill a wide range of area including the channels used by the drone and controller. Eventually, the drone that cannot receive the next command of the controller becomes to continuously carry out the command sent by the previous signals and get out of the user's control.

3.3 Packet Analysis and Acquisition of Control Authority

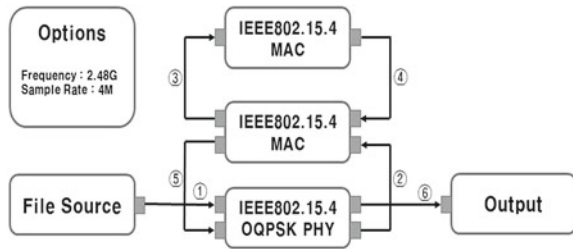
The channels used between the drone and the controller using the channel scanning of Open Sniffer and As shown in Fig. 1, the signals can be collected and analyzed using Wireshark, which is a packet analysis tool. The protocol and important information can be analyzed through the analysis of packet frames and the command part can be guessed using the analysis results. Signals are exchanged using the channel fixed when the drone and controller were initially connected and the information on the channel to be changed is guessed after collecting the packets judging that the information is included in this command. After identifying the fact that the last three bytes of the collected packets are shown to be different from each other as shown in Table 1, the first byte is assumed to be the new channel to replace the current channel and the two latter bytes are assumed to be the CRC. Thereafter, many collected packets are analyzed to identify the channel. Using such information, the channels being used can be identified even when diverse channels are used.

If guessed values for the address, channel, and command parts are obtained utilizing HackRF equipment and the Python open source program GNU Radio, wireless communication can be diversely controlled. The waveforms in the OQPSK mode, which is a ZigBee communication modulation method, can be generated and transmitted using gr-ieee802-14-4 (Aug 7, 2016 Ver) among the open libraries of the GNU Radio. The gr-ieee802-14-4 library is a library fabricated for control of ZigBee protocols and for transmission to the desired packets, the mac.cc in the gr-ieee802-14-4 library should be revised based on the packet data of the actual drone.

Table 1 Collected packets

	Collected packets
1st collection	007eff20a92c2caa70 <u>04 3f 0f</u>
2nd collection	007eff20a92c2caa70 <u>07 e0 ff</u>
3rd collection	007eff20a92c2caa70 <u>03 45 5a</u>
4th collection	007eff20a92c2caa70 <u>05 c0 ff</u>

Fig. 2 Transceiver_OQPSK.grc code flow chart



Since the provided code of `mac.cc` is designated to an arbitrary address and command data, it is revised into the desired packet that has been analyzed. The control signals identified through Open Sniffer and Wireshark can be delivered to the drone through HackRF and the GNU Radio. An advantage of this method is that since commands are made by analyzing collected signals, even other control commands that have not been collected can be identified using the Brute force method on the command data part. This is an attack method that improved the shortcoming of replay attacks that can be enabled by only collected signals and it can find out the control command to achieve complete control.

Desired commands can be made through the library. Frequently used functions have been fabricated as sample codes. If the `transceiver_OQPSK.grc` file among the files provided as sample codes is revised to fit packet transmission as shown in the Figure, the control over the drone can be acquired using desired packets. Figure 2 is a code flow chart of the modified `transceiver_OQPSK.grc` file.

In addition to drones, studies were also conducted with Philips HUE bulbs and Xiaomi smart home kits. When the signals collected to see the internal information of the packet used by Hue are analyzed with Wireshark, packets in forms similar to those of drones can be identified.

In the case of the packets of Hue and the smart home kits, an IEEE 802.15.4 data layer could be identified as with the packet information of drones and the frame and the addresses of the point of departure and the destination could be also identified. However, those pieces of field information such as frame control and sequence numbers were relatively not standardized and continuously changed. Therefore, the control authority could not be obtained using HackRF One tool in methods such as replay attacks. Largely two reasons why standardized packets or network keys did not appear as such can be guessed. First, network keys cannot be identified because they were encrypted. Even though network keys are delivered in the form of plain texts when they are initially delivered to the Coordinator to establish connections, they are immediately encrypted thereafter. Therefore the procedure to find the encrypted keys is not easy. However, in such cases, the keys may be identified using firmware reverse engineering. Second, the network layers of Philips HUE and Xiaomi smart home kits may not use the security mechanism of the ZigBee network layer. Even if the ZigBee security mechanism is not used, the key may be protected by preventing the exposure of network key by using a policy to encrypt the key in the application layer.

4 Analysis and Discussion

In the present paper, simulations were carried out for three different items; drones, lighting systems, and smart home kits. The control over drones was successfully obtained using sniffing equipment because ZigBee follows IEEE standard 802.15.4 and the detailed information, that is, the frame, sequence, and address of the relevant standard could be identified. Eventually, this indicates that even if the security mechanism of ZigBee protocol per se is used even beginners can easily obtain the control authority by utilizing communication sniffing equipment. On the contrary, in the case of Philips HUE bulbs and Xiaomi smart home kits, despite that the packets could be sniffed using ZigBee communication, network key information etc., could not be identified as such information was encrypted because Philips HUE bulbs and Xiaomi smart home kits do not follow the security mechanism of ZigBee but the application programs use independent security mechanisms. Although quite a few threats can be prevented by just preventing those security threats such as key exposure using independent security mechanisms, even when such independent security mechanisms are used, the communication keys being used can be sufficiently exposed by undergoing reverse engineering or memory analysis processes.

5 Conclusion

Past SCADA systems have been blindly trusted to be safer than other systems in terms of security for the reason that they used closed internal networks and non-disclosed communication protocol. However, the simulations carried out in the present paper indicate that at the time when wireless SCADA networks use the standardized protocol termed ZigBee communication, the SCADA networks may become communication networks that are not safe any longer. SCADA networks are exposed to not only the threats of attacks such as replay attacks, sniffing, and control authority acquisition but also those security threats such as jamming that would paralyze the entire system leading to major accidents. Therefore, countermeasures against such security threats should be devised. Although the development of communication technologies greatly contributed to enhancing the level of overall society, the tendency to have interest in only the development of technologies led to the negligence of the reinforcement of information protection resulting in poor responses to incidents. In particular, security incidents occurring in SCADA systems used for control of major national and social infrastructures may cause direct adverse effects. Therefore, related security technologies should be studied and developed to be prepared for such incidents.

Acknowledgments This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00313) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation), the National Research Foundation of Korea (NRF) grant funded by

the Korea government (MSIT) (No. 2018R1D1A1B0705647) and Yeungnam University Research Grant.

References

1. Boyer SA (2009) SCADA: supervisory control and data acquisition. International Society of Automation
2. Labuda DS (2006) Distributed control system. U.S. Patent No. 7,092,768. 15
3. Baker GH, Berg A (2002) Supervisory control and data acquisition (SCADA) systems. The critical infrastructure protection report 1.6
4. Sewio. <http://www.sewio.net/open-sniffer/>
5. Hackrf. <https://greatscottgadgets.com/hackrf/>
6. Gnuradio. <https://www.gnuradio.org/>
7. Syverson P (1994) A taxonomy of replay attacks [cryptographic protocols]. In: Computer security foundations workshop VII, 1994. CSFW 7. Proceedings. IEEE
8. Muraleedharan R, Osadciw LA (2006) Jamming attack detection and countermeasures in wireless sensor network using ant system. Proc SPIE 6248

How Will Blockchain Technology Affect the Future of the Internet?



Geun-Hyung Kim

Abstract Today, the Web is indispensable in daily life. However, the Web has become centralized, because of business models. The centralized Web has brought a few major issues that will likely get worse in the near future. In this paper, we discuss problems of the centralized Web and the necessity of decentralized Web, the potential of blockchain for decentralized web architecture. And then, we also discuss the decentralized web architecture that circumvents Internet gatekeepers like Google and Facebook and takes control of our data back from the giant social media companies.

Keywords Blockchain · Centralized web · Decentralized web · Decentralized Internet · Stateful web

1 Introduction

Many concepts that we have used in blockchain technology were envisioned by Harber and Stornetta [1]. Their work focused on timestamping documents to be able to verify that a document was at a certain time in a certain version, by storing hash values in a timestamped block on the blockchain with which no one could tamper. They also adopted the Merkle tree to enhance efficiency by enabling a single block to include more documents. Satoshi Nakamoto, in 2008, conceptualized the first peer-to-peer version of cryptocurrency with blockchain technology and described in detail how blockchain technology was well equipped to strengthen digital trust in terms of the decentralization aspect in which no trusted intermediary was required [2].

The Internet was originally invented as a decentralized autonomous system in which element technologies, such as inter- and intra-domain routing, DNS, and so on, operate in a distributed fashion. However, the model for applications and

G.-H. Kim (✉)

Game Engineering Major, Dong-eui University, 176, Eomgwangno, Busanjin-gu, Busan 47340, Korea

e-mail: geunkim@deu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_40

several infrastructure services on the Internet has evolved to become more centralized because of business models that depend on centralized accounting and administration [3].

The Web has been one of the representative open application platforms in the Internet since its invention. The Web, which was invented by Tim Berners-Lee, is now 30 years old. The Web connects a great deal of information on the Internet in a hypertext manner, providing a foundation for users to search conveniently. The Web technology basically consists of three main components as follows: the URL (Unified Resource Locator) representing a location to a specific site, HTTP (Hyper Text Transfer Protocol), the protocol for sending and receiving request and response messages, and HTML (Hyper Text Markup Language), the markup language for creating easily hypertext pages. The Web 1.0 and Web 2.0 have revolutionized information and interactions respectively. Right now, relatively few giant social media companies are responsible for hosting essential elements of what we consider the Internet and keeping our e-mail, social media, and web pages available to all. These giant companies that own those hosting servers have a huge impact on how the Internet runs. Current centralized Web platforms leave the Internet open to a few major vulnerabilities that will likely only get worse in the near future, and the least of which is giant companies having power vast amounts of data [4].

In this paper, we discuss problems of the centralized Web and the necessity of decentralized Web, the potential of blockchain for decentralized web architecture. And then, we also discuss the decentralized web architecture that circumvents internet gatekeepers like Google and Facebook and takes control of our data back from the giant social media companies. The remainder of this paper is organized as follows. Section 2 provides an overview of the related works and we discuss the decentralized Web architecture in Sect. 3. Finally, Sect. 4 concludes our study.

2 Related Works

2.1 *Centralized Web Versus Decentralized Web*

The Web was initially designed as a decentralized architecture, but the Web has been significantly centralized since the early 2000s, with the advent of Web 2.0. In the first stage of the Web's evolution (called Web 1.0), content creators were few and the majority of users only acted as a consumer of contents [5]. The open web platform, the collection of open technologies enabling the Web, was emerged by the early 90s and has driven the Web 2.0 era. Any participant in Web 2.0 can be a content creator due to the emerge of new technologies, such as mashups, AJAX (Autonomous JavaScript and XML), and REST (Representational State Transfer) API in the open web platform. The essential characteristics of Web 2.0 are openness, freedom, and collective intelligence by way of user participation [5]. With the advent

of Web 2.0, we began to communicate with each other and to share the information via centralized service platforms provided by giant companies.

As commercial interests grew along with the development of Web technologies, many service platforms, related to the social media, emerged. At the moment, a small number of very powerful platform companies controlled most of the actions on the Internet. These giant companies are known as the FAANGs (Facebook, Amazon, Apple, Netflix, and Google, Microsoft, and Twitter). Over time, the Web has been becoming more centralized in the course of technology. In the centralized Web, the giant companies have monopolized control of user data. The monopoly on data is causing new set of problems that have emerged alongside the consolidation of our online communication [6].

With all our data in the hands of few giant monopolistic companies, we are facing increased hackers, surveillance, censorship, data breaches, misinformation, and so on. For example, Google cooperated with Chinese authorities to run a censored search engine for China, it has been mounting concerns by human right groups about the future of the Web [6]. The Egyptian government blocked around 500 websites as of February 2018. Specially, internet traffic to and from Egypt across 80 internet service providers around the world dropped precipitously on 27–28 January. Thousands of websites, in 2017, had experienced downtime, because of AWS (Amazon Web Service) failure. The source of this massive internet outage is the monopolistic nature of major cloud service providers. With the emergence of centralized service platforms, internet traffic has got centralized as well. Therefore, Internet users are losing their control over the contents they wish to read. It has also compromised the privacy of Internet users and becomes a target for hackers.

Nearly all demographic groups in the US have considered the social media as the most dominant source of news [7]. Facebook is the dominant source of news on government and politics for Millennials especially. From these trends, a few giant platforms can influence significantly over what media users consume on a daily basis. So, these platforms can control what is possible to publish, and they control if others can be likely to discover it [4, 7]. Here are what all users of the centralized Web need to worry about, resulting from risks posed by the centralized Web.

- **Direct Censorship:** Service platforms controlled by the giant companies are more prone to direct censorship and surveillance pressures from the government than decentralized alternatives.
- **Indirect Censorship:** The potential for unintentional or intentional biases equipped in the curation algorithms of social media platform may bring the proliferation of “fake news” or “click-bait headlines”.
- **Abuse of Curatorial Power:** It has been suspected that its employees on Facebook systematically suppressed the discovery of conservative content on their platform.
- **Exclusion:** Giant platforms, like Facebook and Twitter, provided important civic spaces for social and political discourse. So, these spaces should provide unprecedented opportunities for people to reach a global audience and engage in conversions with others around the world. But the reality is not so straightforward.

- Abuse of Privacy and monetize data: Users should share almost personal information with all digital transactions and interactions in order to join social media. Companies use personal information without an agreement to establish the marketing model and to infer user's decisions and make their money off of selling your data to the advertiser who can target you better.
- Single Point of Failure: The service platform becomes a single point of failure for the website as a whole.
- Data breach: Giant companies have often hidden their data breach from consumers for years.

2.2 *Blockchain*

Blockchain technology is a decentralized data storage that stores a registry of assets and transactions across a peer-to-peer network. It is a public registry of who owns and who transacts what. The transactions are secured through cryptography and the transaction history gets locked in blocks of data that are then cryptographically linked together and secured. This creates an immutable, unforgeable record of all of the transactions across this network. The record is replicated on every node in the network. Unlike the existing Internet, blockchain enables users to deliver the value without relying on the third party. Blockchain will be able to delivery various values, economic value like cryptocurrency, stocks, computing resource, real estate, automobile use rights in a shared economic society, and intellectual property rights.

Several cryptographic technologies such hashing and digital signature have been used in the block-chain. Hashing is a method of calculating a relative unique fixed-size output (called digest) for an input of nearly any size (e.g., a video stream, a text file, or an image) and is designed to be one-way and collision-free. Since it results in completely different digests even only single bit in input data has been modified, it provides the integrity of a block data in the blockchain. For digital signature, the asymmetric-key cryptography is utilized to provide the ability to verify the identity of someone who participates in a transaction. Each user possesses a pair of private key and public key. The private key, regarded as the identity and security credential of the user, is used to sign transactions digitally and digitally signed transactions are sent to whole nodes. The public key is used to validate the transactions that are signed with the private key. When a new transaction occurs, user submits a new transaction to the blockchain ledger. A new transaction will be copied and distributed among every node in the blockchain platform and stored in a queue until a mining node adds it to the blockchain by creating a block.

Blockchains allow us to write code, binding contracts, between individuals and then guarantee that those contracts will bear out without a third part enforcer. Blockchain redefines how digital trust mechanisms through distributed consensus mechanisms and transparent temper-evident record-keeping.

3 Decentralized Web Architecture

Fundamentally, the decentralization Web is about enabling choice, by breaking up artificially coupled decisions into individual options that can be combined at one's pleasure [8]. From the concept of decentralized Web, we should be able to interact with web sites and other people without commitment to single social media platform. In terms of taking back control of our personal data, decoupling the sensitive personal data from services. This allows users to be able to enjoy the applications they want and to store data where they specify. We can select any service provider to store our text, photos, and videos to store them on our own Web storage and depend on any third-party service to interact with data, regardless of storage location. As an example, the identity data for crucial identity service can be provided by the Web storage.

Therefore, we can place annotations, and comments on anything we want, without fear of them being censored or deleted, because we do not require any permission to publish data in our own data storage. To guarantee highly granular access to personal data, user gives permission selectively to friends or application to access specific parts of their data. These permissions can be changed and revoked at any time.

Decentralized Web requires the nature of applications to evolve from silo architecture to shared architecture. In terms of MVC (model, view, and control) design pattern, centralized Web combines view and model. On the other hand, decentralized Web handles the model (data) separately from view and control (service logic). As described earlier, each data can be accessed with specific permission rights. When granted specific access rights, annotations, photo, music, video etc. uploaded into our data storage by application can be accessed by social media applications. Events in my personal calendar that have public visibility can be shown up the same social media applications. Our social friends can view the parts of our data to which we grant them access through whatever application they want to use. The applications in the centralized Web compete in a single market, based on data ownership. So, new innovative competitors in the centralized Web may have a trouble entering market because of lack of data of customers.

The decentralized Web (called Web 3.0) has the potential to transform our experiences into the new world. In the decentralized Web, services are distributed rather than localized, when users own and control their data and where small players take back powers for giants like Google and Amazon. The Web 3.0 has potential to revolutionize contracts and value exchanges. It changes the data structures in internet backbone introducing a universal state layer, open by incentivizing network actors [9].

Today, the internet we use does not have a mechanism to transfer the status of who is who, who owns what, and who has the right to do what. However, the state is a key property for managing values. Blockchain introduced a method for each participant in a network to hold and transfer value in a digitally native format, without the need for trusted intermediaries. The consensus protocol is designed in a way that the network can collectively remember preceding events or user interactions. The blockchain

protocol can be seen as a game-changer, paving the way to a more decentralized Web.

4 Conclusion

More and more of user-generated content is now hosted on centralized servers, belong to a small group of giant companies. This trend has brought the centralized Web that has several issues, not the least of which is giant companies having power vast amounts of data. The Web 3.0 has the potentials to revolutionize contracts and value exchanges and to decouple data and related applications. In this paper, we discussed problems of the centralized Web, the necessity of decentralized Web, and the potential of blockchain for decentralized web architecture. And then, we also discussed the decentralized web architecture that circumvents internet gatekeepers like Google and Facebook and takes control of our data back from the giant tech companies.

Acknowledgements This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MOE: Ministry of Education): (NRF-2017R1D1A1B03035074).

References

1. Haber S, Stornetta WS (1991) How to time-stamp a digital document. *J Cryptol* 3(2):99–111
2. Nakamoto S (2009) Bitcoin: a peer-to-peer electronic cash system. Cryptography Mailing list at <https://metzdowd.com>
3. Charter for Research Group—Decentralized Internet Infrastructure Research Group (DINRG). <https://datatracker.ietf.org/group/dingr/about>. Accessed 20 Oct 2019
4. Rowe A (2018) Everything you need to know about the decentralized Internet, Article of TechCo. <https://tech.co/news/decentralized-internet-guide-2018-02>. Accessed 20 Aug 2019
5. Cormode G, Krishnamurthy B (2008) Key difference between Web 1.0 and Web 2.0. *First Monday* 3(6)
6. Corbyn Z (2018) Decentralisation: the next big step for the world wide web, Article of The Guardian. <https://theguardian.com/technology/2018/sep/08/centralisation-next-big-step-for-the-world-wide-web-dweb-data-internet-censorship-brewster-kahle-2018-09>. Accessed 20 Aug 2019
7. Barabas C, Narula N, Zuckerman E (2017) Defending Internet freedom through decentralization: back to the future? Report of MIT digital currency initiative and the center for civic media
8. Verborgh R (2019) Re-decentralizing the web. In: Seneviratne O, Hendler J (eds) *Linking the world's information: Tim Berners-Lee's invention of the world wide web*. ACM (accepted for publication)
9. Voshmgir S (2019) Tokenized networks: web 3, the stateful web. *Token Economy*

An Implementation of DAQ System for a Smart Fish Farm: Based on a Semi Circulation Filtration System in S. Korea



Joo H. Jean, Na E. Lee, Yoon H. Lee, Jea M. Jang, Moon G. Joo, Byung H. Yoo, and Jea D. Yoo

Abstract We implemented a data acquisition system for an automated system for smart fish farms. The fish farm is located in Jang Hang, S. Korea, and designed using a circulating filtration system. Information from aquaculture pools is automatically measured by pH sensors, DO sensors, and water temperature sensors and stored in a server database. Collected data using Modbus protocol are used to optimize pool water quality, to predict the rate of growth of the fish, and to deliver food automatically as planned by fish farm. Collected data are delivered to the user's PC for analysis and monitoring and to mobile phone by using JSON protocol. The developed automation system allows fish farmers to improve fish productivity and maximize profits.

Keywords DAQ system · Fish farm · Circulation filtration system · Database · LabVIEW

1 Introduction

With the development of the fourth industrial revolution using ICT technology, big data technology, and networks, automation of existing aquaculture is proceeding rapidly [1–3]. This is called a smart fish farm. The purpose of smart fish farms is to systematically manage the breeding environment of fish to improve productivity and increase the income of fish farms.

Fish farmers select fish species that are resistant to various fish diseases and have good food efficiency and carry out large-scale farming in pools on the ground. In order to maximize the cultivation efficiency of selected fish species, it is necessary to create and maintain the environment of the farms optimally.

J. H. Jean · N. E. Lee · Y. H. Lee · J. M. Jang · M. G. Joo (✉)
Department of Information and Communications Engineering, Pukyong National University,
Busan, South Korea
e-mail: gabi@pknu.ac.kr

B. H. Yoo · J. D. Yoo
Jang Hang Fish Farm, Chungcheongnam-do, South Korea

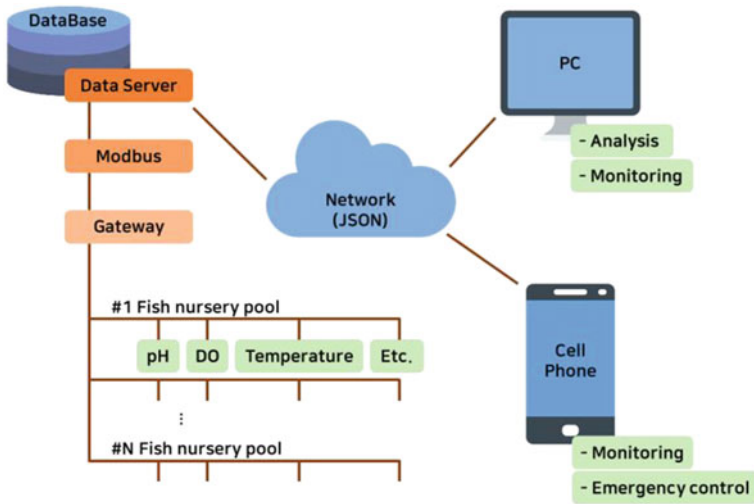


Fig. 1 Overview of the data acquisition (DAQ) system

Typically, it is necessary to automatically collect oxygen concentration, water temperature, and pH concentration, and analyze them to maintain constant values. NH_3 concentrations, NO_2 concentrations, feed rates, and fish's average weight, which are not automatically analyzed or require time for analysis, should be added later by the user to form a complete database.

Figure 1 shows an overview of the DAQ (Data acquisition) system of the Jang Hang smart fish farm in Korea based on the semi the circulation filtration system, which is an environmentally friendly fish farm. The purpose of the DAQ system is to monitor water quality and to predict fish growth rate for optimal fish farming.

Information from each aquaculture pool is automatically collected from pH sensors, DO sensors, and water temperature sensors and stored in a server database. The communication protocol between the sensors and the data server uses the international standard, Modbus. Data that is not automatically collected are entered manually by the user to construct a database.

Fish farmers use data transmitted from data servers to on-site PCs to monitor fish farm conditions and analyze data to optimize fish growth. The system was developed to enable fish farmers to monitor data on their mobile phones, with the added ability to alert and control emergencies. JSON was used as a communication protocol between the data server and the PC or mobile phone.

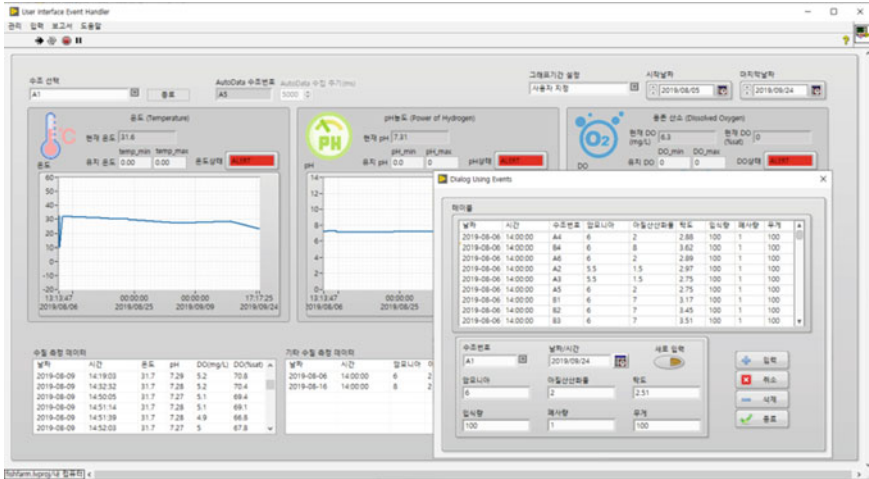


Fig. 2 DB management tool for a smart fish farm

2 Developed Tools for Smart Fish Farm

2.1 Fish Farm DB Management Tool

Information from each aquaculture pool is automatically collected from pH sensors, DO sensors, and temperature sensors and stored in a server database. Data that is not automatically collected is entered manually by the user to construct a database. Fish farmers can view the data history of each farm as shown in Fig. 2.

This program is written using LabVIEW and uses the MS access database. Data are automatically collected every 60 s. User can print out an Excel data sheet for each pool.

2.2 Simulation Tool for Predicting Fish Growth

A software tool is developed to simulate the growth of target fish when aquaculture environment variables are controlled. The fish growth model used in the simulation was developed by Ursin [4], modified by Bolte [5], and continues to be developed. The fish growth calculation is expressed using the difference between fish anabolism and catabolism as shown

$$\frac{dW}{dt} = HW^m - kW^n,$$

where

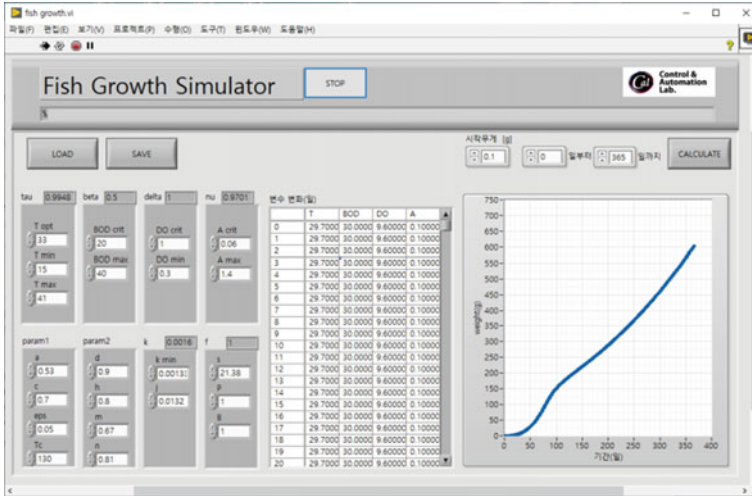


Fig. 3 Simulation of a fish growth

W : fish weight (g)

t : time (day)

H : coefficient of anabolism

M : index of anabolism

k : coefficient of catabolism

n : index of catabolism.

The user can check the growth rate of the fish by changing the temperature of the tank, BOD, dissolved oxygen and ammonia on a daily basis as shown in Fig. 3.

This program is written in LabVIEW. The solution of the differential equation is calculated by the fourth-order Runge-Kutta equation.

2.3 Automatic Fish Feeder Management

As shown in Fig. 4, a tool for automatically feeding fish for each aquaculture pool is designed. The amount of food required is automatically calculated based on water temperature, fish weight and fish population in a pool. The user can check when the automatic feeding machine is running out of feed or there is a problem with the machine.

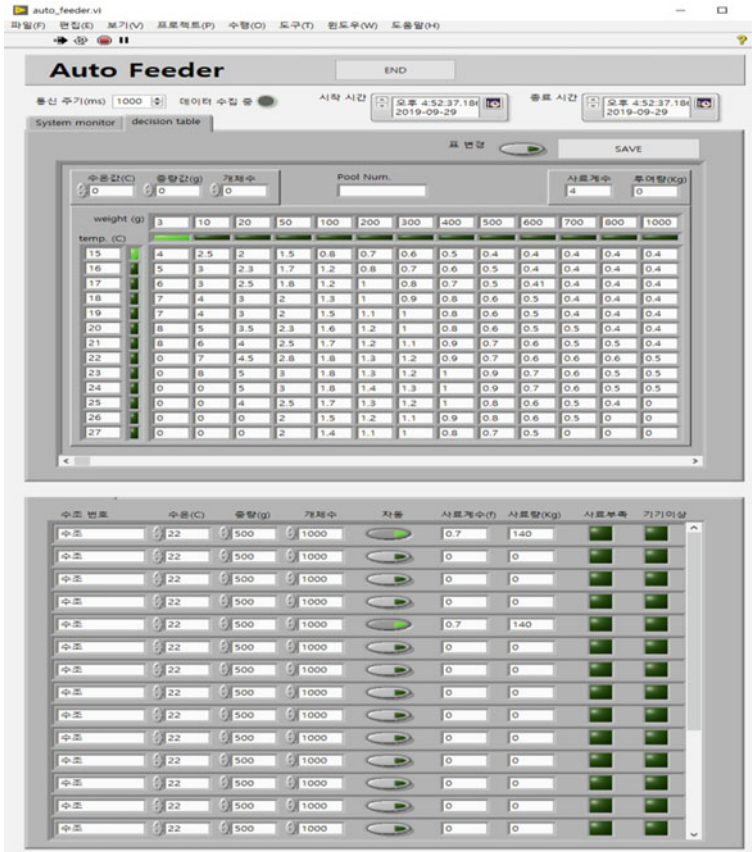


Fig. 4 Decision table and System monitor of auto feeder

3 Conclusion

We implemented a data acquisition system for an automated system for smart fish farms. Collected data using Modbus protocol are used to optimize pool water quality, to predict the rate of growth of the fish, and to deliver food automatically as planned by fish farm. Collected data are delivered to the user's PC for analysis and monitoring and to mobile phone by using JSON protocol.

The system is programmed using LabVIEW and includes DB management tool, fish growth simulation, and automatic fish feeder. The developed automation system allows fish farmers to improve fish productivity and maximize profits.

Acknowledgements This research was supported by the Korea Institute of Marine Science and Technology (KIMST) (Grant number: 20180352).

References

1. Chen JH, Sung WT, Lin GY (2015) Automated monitoring system for the fish farm aquaculture environment. In: IEEE international conferences on systems, man, and cybernetics
2. Huh JH (2016) Design and android application for monitoring system using PLC for ICT-integrated fish farm: advanced multimedia and ubiquitous engineering. LNEE 393:617–625
3. Ullah I, Kim D (2018) An optimization scheme for water pump control in smart fish farm with efficient energy consumption. Processes 6(65)
4. Ursin E (2011) A mathematical model of some aspects of fish growth, respiration, and mortality. J Fish Res Board Can 24:2355–2453 (2011)
5. Bolte JP, Nath SS, Ernst DH (1995) A decision support system for pond aquaculture, In: Twelfth annual administrative report. Pond Dynamics/Aquaculture CRSP Office of International Research and Development, p 95

Design of Middleware to Support Auto-scaling in Docker-Based Multi Host Environment



Minsu Chae, Sangwook Han, and Hwa Min Lee

Abstract With the spread of smart devices, the use of big data, and the proliferation of the Internet of Things, virtualization technology for cloud servers have become important worldwide. Also, research has been conducted to efficiently manage the resources of hosts in VMs. Container-based virtualization has less performance degradation than VMs because there is no emulation for the operating system. Using the Docker API is slow to measure. In this paper, we implement the resource measurement module of Job nodes and design middleware that supports auto-scaling in auto-scaling module and Docker-based multi-host environment.

Keywords Docker · Cloud · Auto-scaling · Multi host

1 Introduction

With the spread of smart devices, the use of big data, and the proliferation of the Internet of Things, virtualization technology for cloud servers have become important worldwide. In addition, research has been conducted to efficiently manage the resources of hosts in Virtual Machines (VMs) [1, 2]. However, in the case of a VM, the VM emulates an operating system [3], so there is performance degradation of the VM [4, 5]. Accordingly, container-based virtualization technology has been studied [6, 7]. Container-based virtualization has less performance degradation compared to VM because there is no emulation for the operating system [8] Because container-based virtualization does not emulate the operating system, only the resource usage

M. Chae · S. Han

Department of Computer Science, Soonchunhyang University, Asan, South Korea
e-mail: cms@sch.ac.kr

S. Han

e-mail: sanguk@sch.ac.kr

H. M. Lee (✉)

Department of Computer Software Engineering, Soonchunhyang University, Asan, South Korea
e-mail: leehm@sch.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_42

301

of the running process is required [6–8]. In the case of VM, resource utilization is low, because it allocates the requested resource even if do not use resource that requested by the user [8]. Also, the VM has a disadvantage that it takes about 1 min during the bootstrap process. In the case of a container, it takes a few seconds. Accordingly, this paper designs middleware that supports auto-scaling in Docker-based multi-host environment.

2 Related Works

2.1 Docker

Docker is a container-based application automation open source project. Docker used the Linux container (LXC) in the past, recently Docker uses namespaces and cgroups directly [6, 8]. Docker uses images that are needed to run programs, such as environment files, runtime libraries, and system libraries. In other words, Docker uses an image to bundle and distribute the required packages for a process. Previously it was only available on Linux, Recently it available on Windows 10 Professional through Hyper-V [9, 10].

2.2 Auto-scaling

In April 2008, Animoto increase VMs due to a sudden increase in traffic [11]. And traffic dropped sharply to normal levels [11]. So, Animoto reduced the number of VMs [11]. Because of the sudden increase in traffic, auto-scaling has emerged as one of the key features in cloud environments. The typical cloud providers that provide autoscaling are Amazon Web Service (AWS), Google Cloud Platform (GCP), and MS Azure. Several studies [12–14] have been conducted on auto-scaling.

2.3 CoreOS

CoreOS is the operating system for Docker [15]. CoreOS supports managing multiple hosts [15]. In particular, CoreOS determines the host by the priority specified when creating the container. It also supports creating multiple containers with the same image at once [15]. However, there is a disadvantage that does not support auto-scaling. Accordingly, the administrator must manage the containers.

2.4 Kubernetes

Kubernetes is a container management tool designed by Google [16]. Kubernetes supports autoscaling when configured [16]. However, volume sharing issues remain when using auto-scaling. If the volume sharing setting is set to hostPath, there is a disadvantage that data cannot be obtained from the previous host when executed on another host [17, 18]. If the volume is set to gitRepo, a site based on Contents Management System (CMS) has a problem of uploading a version to the git repository every time an attachment is uploaded if the file in the post is uploadable. Otherwise, there is a problem that makes uploading attachments impossible [17, 18]. If the volume setting is set to PersistentVolume, there is an advantage that a container created on another host can be shared and used, but there may be a delay due to network communication [17, 18].

3 Middleware Design

3.1 System Architecture

Figure 1 shows the architecture of the middleware proposed in this paper. Job nodes periodically send container usage to Master nodes. The master node performs auto-scaling when the CPU usage of the container running on the job node is higher than the threshold.

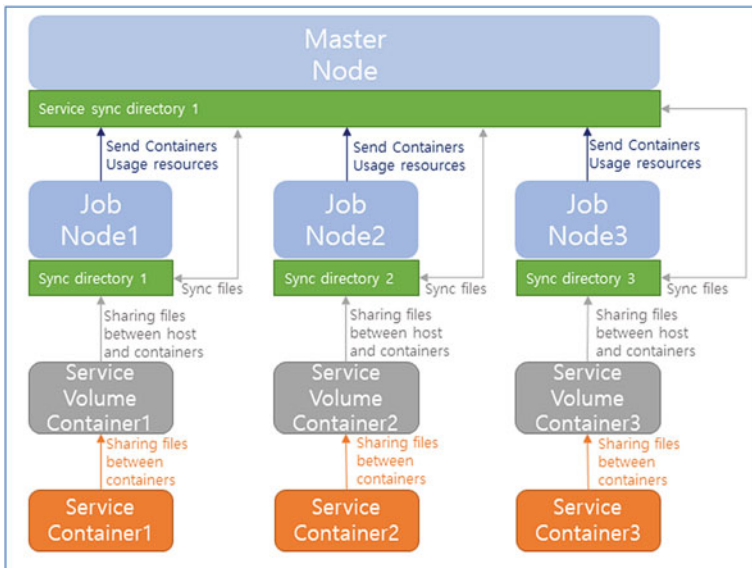


Fig. 1 The architecture of proposed middleware

3.2 Resource Measurement Module

If the resource measurement module uses the Docker API, it can request resource usage measurement of one container at a time. Therefore, in this paper, we analyzed the internal file of `cgroup` which is used to isolate the process in Docker container. There is a directory named container ID in the `/run/docker/runtime-runc/moby` directory. A `state.json` file exists under the directory named container ID. We confirm the directories path in use inside the file.

Figure 2 shows the contents of the `state.json` file in the `/run/docker/runtime-runc/moby` directory. Inside the value of the `cgroups_paths` key, the directory path used internally is stored. In the case of CPU, the corresponding virtualized CPU resource amount and the total amount of virtualized CPU can be checked in the corresponding directory. It is possible to check the utilization rate of each core, but the utilization rate of all cores is unknown. Therefore, we did not parse the files under the CPU directory.

In the case of `pids`, there is a “tasks” file under that directory, and inside that file there is a list of running `pids`. In the case of Docker, it is isolated and running on the host, so we can use the `Top` command to check the CPU usage in use. The `Top` command can check CPU usage and running process name. CPU usage in the `Top` command is expressed as a percentage, and 100% is the sum of the utilization running on each core, not the total utilization of the host’s CPU. For example, suppose that the CPU utilization is 200%. If CPU has four cores, the average usage per core can be calculated as 50%. However, the `Top` command does not provide a utilization rate

Fig. 2 Sample of the `state.json`

```
{
  "id": "1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
  "config": {
    "no_pivot_root": false,
    "parent_death_signal": 0,
    "ounts": [
      {
        "source": "proc",
        "destination": "/proc",
        "device": "cgroups",
        "path": "/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
        "cpu_shares": 0,
        "cpu_quota": 0,
        "cpu_period": 0,
        "cpu_rt_quota": 0
      }
    ],
    "cgroup_paths": {
      "blkio": "/sys/fs/cgroup/blkio/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "cpu": "/sys/fs/cgroup/cpu,cpuacct/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "cpuacct": "/sys/fs/cgroup/cpu,cpuacct/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "cpuset": "/sys/fs/cgroup/cpuset/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "devices": "/sys/fs/cgroup/devices/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "freezer": "/sys/fs/cgroup/freezer/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "hugetlb": "/sys/fs/cgroup/hugetlb/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "memory": "/sys/fs/cgroup/memory/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "name=systemd": "/sys/fs/cgroup/systemd/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "net_cls": "/sys/fs/cgroup/net_cls,net_prio/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "net_prio": "/sys/fs/cgroup/net_cls,net_prio/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "perf_event": "/sys/fs/cgroup/perf_event/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07",
      "pids": "/sys/fs/cgroup/pids/docker/1d96a2b92e2d79aed5caf411c5f9d54b4cc46f967c51afe07"
    }
  },
  "namespace_paths": {
    "NEWIPC": "/proc/19067/ns/ipc",
    "NEWNET": "/proc/19067/ns/net",
    "NEWNS": "/proc/19067/ns/mnt",
    "NEWPID": "/proc/19067/ns/pid",
    "NEWUSER": "/proc/19067/ns/user",
    "NEWUTS": "/proc/19067/ns/uts"
  },
  "external_descriptors": [
    "pipe: [680372259]",
    "pipe: [680372259]",
    "intel_rdt_path": ""
  ]
}
```

for each core, so it uses its own value rather than dividing it by the number of cores. Thus, we can accurately determine the workload of each container.

Algorithm 1. The resource measurement module

```

program measure () {
  containerIds = os.list("/run/docker/runtime-runc/moby")
  resources = {}
  for containerId in containerIds
    resources[containerId]['memory'] = parseMemory(containerId.memory)
    resources[containerId]['cpu'] = parseCPUviaPidFiles(containerId.pids)
  return resources
}

```

3.3 *Auto-scaling Module*

The auto-scaling module monitors the resource utilization of each container obtained through the resource measurement module. If the CPU threshold for auto-scaling specified for the service is exceeded, one of the job nodes is selected to create a container. If the created container has CPU usage below the threshold, it will be automatically deleted.

Algorithm 2. The auto-scaling module

```

program AutoScaling () {
  containers = measure()
  for container in containers
    if container['CPU'] >= thresholdLimitUp[container['ServiceName']]
      containerCreate(container.image)
    else if container['CPU'] <= thresholdLimitDown[container['ServiceName']]
      container['check']++
    else if container['CPU'] > thresholdLimitDown[container['ServiceName']]
      container['check']=0
    if container['check'] >= thresholdCheck[container['ServiceName']]
      containerRemove(container)
}

```

3.4 *Storage Sharing Module*

In the case of auto-scaling on multiple hosts, it is important that the container must run even if the container is created on another host. Accordingly, the master node

sets directories to be shared for each service, and the job node and the master node synchronize the directories to be shared.

Algorithm 3. The storage sharing module

```
program sharing () {
  while(true)
    if user.isChangeFiles()
      files = user.getChangeFiles
      for file in files
        file.sync()
        sendAllwithoutUser(user, file)
}
```

4 Conclusion

In this paper, we designed middleware that supports auto-scaling in Docker-based multi-host environment. The designed middleware has the following effects. First, it provides load balancing for each service. This makes it possible to provide a smooth service even if all the services are suddenly crowded. Second, resource efficiency is improved by closing and deleting containers that do not have CPU utilization. Third, using Docker provides optimal auto-scaling using a limited amount of resources. In the future, it is expected to apply this to Edge Server to provide low service latency in the 5G environment and to use minimal resources. We will implement the middleware we designed.

Acknowledgements This research supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00720 & IITP-2019-2015-0-00403) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

References

1. Graubner P, Schmidt M, Freisleben B (2011) Energy-efficient management of virtual machines in eucalyptus. In: 2011 IEEE 4th international conference on cloud computing. IEEE
2. Leelipushpam P, Jeba G, Sharmila J (2013) Live VM migration techniques in cloud environment—a survey. In: 2013 IEEE conference on information & communication technologies. IEEE
3. Ye K et al (2010) Analyzing and modeling the performance in Xen-based virtual cluster environment. In: 2010 IEEE 12th international conference on high performance computing and communications (HPCC). IEEE
4. Morabito R (2016) A performance evaluation of container technologies on Internet of Things devices. In: 2016 IEEE conference on computer communications workshops (INFOCOM WKSHPs). IEEE

5. Barham P et al (2003) Xen and the art of virtualization. *ACM SIGOPS Oper Syst Rev* 37(5)
6. Rosen R (2014) Linux containers and the future cloud. *Linux J* 240(4):86–95
7. Tosatto A, Ruiiu P, Attanasio A (2015) Container-based orchestration in cloud: state of the art and challenges. In: 2015 Ninth international conference on complex, intelligent, and software intensive systems. IEEE
8. Chae MS, Lee HM, Lee K (2019) A performance comparison of linux containers and virtual machines using Docker and KVM. *Clust Comput* 22(1):1765–1775
9. Fink J (2014) Docker: a software as a service, operating system-level virtualization framework. *Code4Lib J* 25
10. <https://www.docker.com>
11. <https://www.flexera.com/blog/cloud/2008/04/animotos-facebook-scale-up/>
12. Jiang J et al (2013) Optimal cloud resource auto-scaling for web applications. In: 2013 13th IEEE/ACM International symposium on cluster, cloud, and grid computing. IEEE
13. Mao M, Humphrey M (2011) Auto-scaling to minimize cost and meet application deadlines in cloud workflows. In: SC'11: Proceedings of 2011 international conference for high performance computing, networking, storage and analysis. IEEE
14. Hasan MZ et al (2012) Integrated and autonomic cloud resource scaling. In: 2012 IEEE Network operations and management symposium. IEEE
15. <https://coreos.com/>
16. <https://kubernetes.io>
17. <https://kubernetes.io/docs/user-guide/volumes>
18. Heidari P, Lemieux Y, Shami A (2016) Qos assurance with light virtualization-a survey. In: 2016 IEEE international conference on cloud computing technology and science (CloudCom). IEEE

Gated Convolutional Neural Networks for Text Classification



Jin Sun, Rize Jin, Xiaohan Ma, Joon-young Park, Kyung-ah Sohn, and Tae-sun Chung

Abstract The popular approach for several natural language processing tasks involves deep neural networks, and in particular, recurrent neural networks (RNNs) and convolutional neural networks (CNNs). While RNNs can capture the dependency in a sequence of arbitrary length, CNNs are suitable for extracting position-invariant features. In this study, a state-of-the-art CNN model incorporating a gate mechanism that is typically used in RNNs, is adapted to text classification tasks. The incorporated gate mechanism allows the CNNs to better select which features or words are relevant for predicting the corresponding class. Through experiments on various large datasets, it was found that the introduction of a gate mechanism into CNNs can improve the accuracy of text classification tasks such as sentiment classification, topic classification, and news categorization.

Keywords Gate mechanism · Convolutional neural networks · Text classification

J. Sun · X. Ma · K. Sohn · T. Chung (✉)
Computer Engineering, Ajou University, Suwon-si, Gyeonggi-do 16499, Korea
e-mail: tschung@ajou.ac.kr

J. Sun
e-mail: jingsun@ajou.ac.kr

X. Ma
e-mail: maxiaohan@ajou.ac.kr

K. Sohn
e-mail: kasohn@ajou.ac.kr

R. Jin · J. Park
School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300160, China
e-mail: jinrize@tjpu.edu.cn

J. Park
e-mail: pjy2018@tjpu.edu.cn

1 Introduction

In natural language processing (NLP), text classification is the process of assigning a piece of text to one or more classes. In recent years, as neural networks show strong power in computer vision [1] and speech recognition [2], neural networks models also show significant power on NLP tasks, such as recurrent neural networks (RNNs) [3, 4] and convolutional neural networks (CNNs) [5, 6]. In neural-network-based models for text classification, words or characters are usually mapped into a lower dimensional space and then regulated by the parameters in neural networks to generate a more compact representation, which is the input of a classifier. Through hierarchical or recursive computation over the input text, neural networks can extract high level features automatically and directly from raw data, thus enabling them to capture global dependencies among words in a sentence and to have a much higher efficiency in processing big data as compared to traditional approaches.

In this paper, we adapt the gated CNN architecture of Dauphin et al. [7] to text classification tasks. In the proposed model, gates are inserted between subsequent CNN layers, thus controlling what information can pass through it, which is considered to improve the efficiency of the networks in finding the important features. To the best of our knowledge, this is the first study in which gated CNNs are applied to a text classification task.

The contributions of this work are as follows:

- We first adapt the gated CNN architecture to the text classification task. Some popular techniques for further improving the neural networks performance, such as residual connection [8] is utilized and analyzed.
- We analyzed some of state-of-the-art neural networks based architectures of text classification on the same datasets for comparison.
- Experiments on four large scale datasets demonstrate the effectiveness of the gate mechanism in CNNs for text classification task.

2 Gated Convolutional Neural Networks for Text Classification

In this study, we present a variation of the gated CNN architecture proposed by Dauphin et al. [7] for text classification tasks. The gates are introduced in CNNs to better control the information of the previous layer’s outputs that should be propagated to the subsequent layers. Within each gated CNN layer, a “gate” operation that is implemented by a sigmoid function is applied after each convolution operation. The dotted box of Fig. 1 illustrates the gate mechanism in CNNs. Given an input $X \in \mathbb{R}^{N^m}$, we define the computed output of a gated convolution layer as follows:

$$Y = (X \cdot W + b) \otimes \sigma(X \cdot V + c) = A \otimes \sigma(B) \quad (1)$$

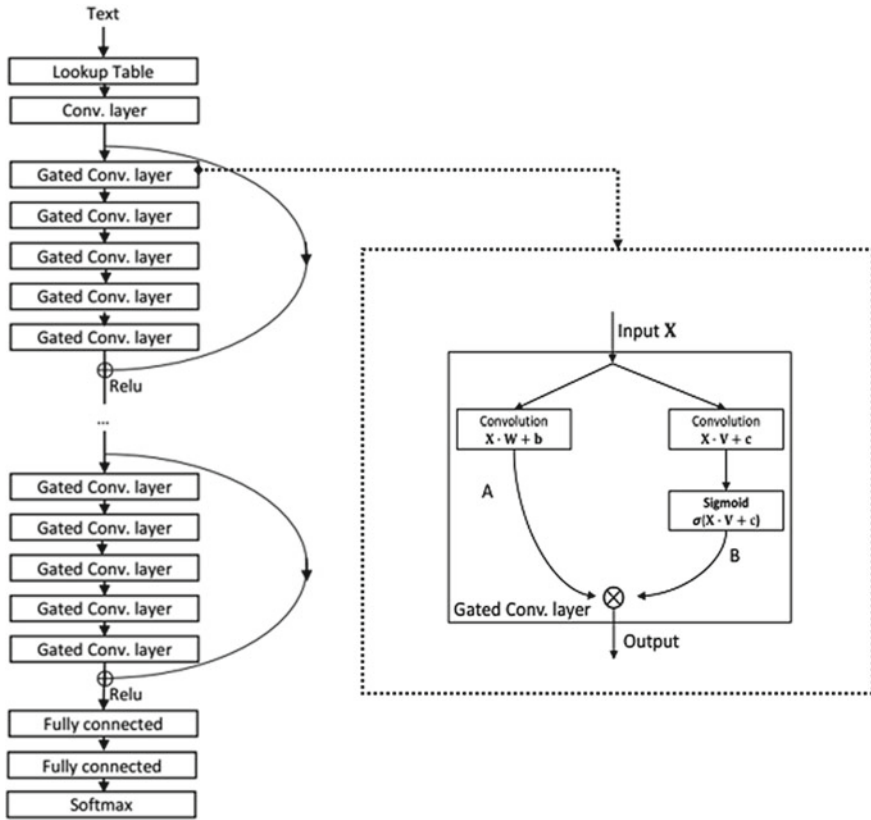


Fig. 1 The architecture of the gated convolutional neural networks for text classification. The gated convolution layer is shown in the right dotted box

where $W, V \in \mathbb{R}^{km}$ are the convolution filters of the layer of size $k \times m$, $b, c \in \mathbb{R}$ are the bias terms. The input of a gated CNN layer follows one of the two paths shown in the dotted box of Fig. 1. One path comprises a convolution operation ($X \cdot W + b$) to obtain a result, which is denoted by A. The other path comprises a different convolution operation given by ($X \cdot V + c$), the result of which is denoted by B, after which a sigmoid function is applied to it to generate values from 0 to 1 as gate. The results of the first path A are then modulated by the gate $\sigma(B)$.

The overall architecture of our model is illustrated in Fig. 1. First, the input text is converted into vectors by using a lookup table. The first layer is a normal convolutional layer without a gate, which is followed by several gated convolutional layers. Every five layers form a residual block [8], and there exists a residual connection from the input to the output of the block. Architectures of various depths are obtained by adding or removing such residual blocks. Finally, a classifier comprising two fully connected layers with a *SoftMax* function generates the final probability vector.

During the back-propagation process, the gradient of the gate is computed as follows:

$$\nabla[A \otimes \sigma(B)] = A' \otimes \sigma(B) + A' \otimes \sigma'(B)B' \quad (2)$$

It should be noted that the first part $A' \otimes \sigma(B)$ provides a linear path for the gradients to back propagate without downscale, thus solving the gradient vanishing problem.

The main adaptation of our model from Dauphin et al. [7] is the activation function, i.e., the classifier that is applied after the fully connected layers. The architecture in the work of Dauphin et al. [7] made use of adaptive *SoftMax* [9], in order to improve the computation efficiency. In our work, the task is to classify sentences, and with only 2 to 14 probabilities, the number of classes is required to be calculated at each iteration. Therefore, we use the common *SoftMax* function as the classifier:

$$p(y = k|X) = \frac{\exp(x \cdot w_k + b_k)}{\sum_{k'}^K \exp(x \cdot w_{k'} + b_{k'})} \quad (3)$$

Here we assume that there are K classes. w_k is the weight in the last fully connected layer, and b_k is the bias. In addition, the hyper parameters in the networks are adjusted to better fit the text classification task.

3 Experiments

3.1 Datasets

We test our model on four large-scale text classification datasets introduced by Zhang et al. [5]: DBPedia, AG's news, Yelp review polarity, and Yelp review full. DBPedia is extracted from the English edition of Wikipedia and consists of 14 non-overlapping classes. The size of the training set is 560,000 and that of the test set is 70,000. AG's news is obtained from AG's corpus of news articles distributed in four classes. The size of the training set is 120,000 while that of the test set is 7,600. Yelp review is obtained from the 2015 Yelp Dataset Challenge. Yelp review polarity is sampled in two classes, positive and negative, with a training set size of 560,000 and test set size of 38,000. Yelp review full is sampled in five classes from 1 star to 5 star, with 650,000 samples for the training set and 50,000 samples for the test set. For more details refer to Zhang et al. [5]. Table 1 shows a summary of these datasets.

Table 1 Summary of datasets

Datasets	#Classes	#Train	#Test	Average length
DBPedia	14	560,000	70,000	53
AG. news	9	4,000,000	790,000	42
Yelp polarity	2	560,000	38,000	147
Yelp full	5	650,000	50,000	146

3.2 Results and Discussions

We evaluate our model using several configurations and list the error rate for all the datasets, while the configurations are set as 15 gated CNN layers with a residual connection inserted and batch normalization applied after each convolution operation. Table 2 shows a comparison of the error rate of various architectures on the datasets. The very deep CNNs [6] obtained the best performance when the architecture is configured at 29 convolutional layers. The results of the proposed model are competitive with those of the hybrid CNN–RNN model of Xiao et al. [10] and outperform both the character-level and word-level CNNs from Zhang et al. [5], except for the result obtained for the Yelp Polarity dataset as compared with word-level CNNs. Despite a slight loss on the DBPedia dataset, our model also performs better than the common LSTM approach.

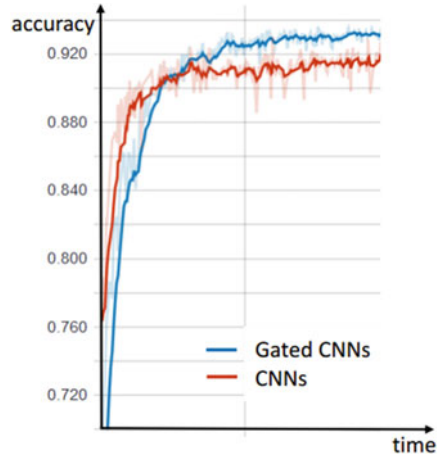
Datasets matters. From the results of all the architectures we listed, the classification task is performed best on the DBPedia dataset, which is an ontology classification dataset. The worst results are generated on the Yelp Full dataset, which is a full level movie review dataset. We can conclude from such results that the neural networks can capture much more distinctive features from the DBPedia text data than from the other datasets. Therefore, how to capture more efficient features specific to a certain dataset is a key issue for performance improvement in text classification tasks. Next, we investigate how to better represent text from a specific dataset.

Effect of gate mechanism in CNNs. Figure 2 shows the effect of the gate mechanism applied to CNNs. From the plot, the architecture with gates has a relatively slow convergence speed at the beginning of the training process, and its accuracy increases

Table 2 Results of our model compared to the state-of-the-art architectures. The error rates from Zhang et al. [5] are the results of models with a small feature size and without data augmentation

Model	DBP.	AG. N.	Yelp P.	Yelp F.
LSTM [5]	1.45	13.94	5.26	41.83
Character-level CNNs [5]	1.98	15.65	6.53	40.84
Word-level CNNs [5]	1.71	11.35	5.56	42.13
CNN-RNN [10]	1.43	8.64	5.51	38.18
Very deep CNNs [6]	1.29	8.67	4.28	35.28
Gated CNNs	1.49	9.27	6.25	38.58

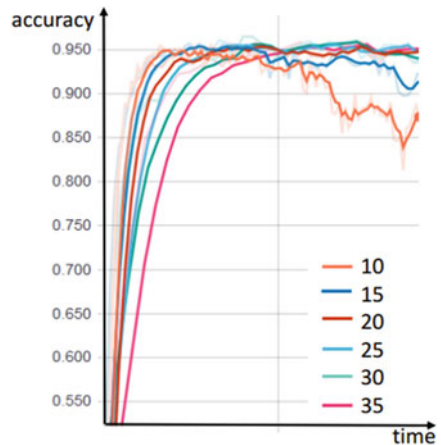
Fig. 2 Effect of gate mechanism in CNNs



as the training continues on. This trend continues until the end of the training process. In contrast, the architecture without gates has a faster convergent speed but results in a lower final accuracy. These results show that the addition of the gates results in the increase of the number of parameters in the networks, thus makes the convergence slower. However, as the training proceeds, the gates can help CNNs achieve a higher accuracy.

Depth of networks. We test the performance of our model with a residual connection inserted in the gated convolutional layers for different depths on the DBPedia dataset. Figure 3 shows the accuracy tendency over time when evaluated on the DBPedia development set during the training phase. From 10 gated convolutional layers to 35 layers, the convergent speed becomes increasingly slower as there are more parameters to train. However, the best accuracies for the various depths for the test set are almost the same with a variance of within 0.1%. However, in the work of

Fig. 3 Comparison of various depths on DBPedia



Conneau et al. [6], their results showed that as the depth increases from 9 to 29 with a residual connection, the test error decreases for all the data sets. In the future, we intend to further explore the effect of increasing the layers with a residual connection.

4 Conclusions

This paper presents a variant of gated CNN architecture for text classification tasks. The architecture is evaluated using several large-scale datasets including ontology classification, sentiment classification, and news classification. The experimental results show that the gate mechanism introduced in CNNs results in a higher accuracy on text classification tasks comparing with some state-of-the-art neural network architectures on the same datasets. In addition, we explore the influence of the networks' depth with a residual connection and the impact of the use of a batch normalization mechanism. Our future work will include the exploration of the applicability of gated CNN architecture on the domain of information retrieval and recommendation systems.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1F1A1058548).

References

1. Siggmony K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
2. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag* 29(6):82–97
3. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence
4. Liu P, Qiu X, Huang X (2016) Recurrent neural network for text classification with multi-task learning. [arXiv:1605.05101](https://arxiv.org/abs/1605.05101)
5. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: *Advances in neural information processing systems*, pp 649–657
6. Conneau A, Schwenk H, Barrault L, Lecun Y (2017) Very deep convolutional networks for text classification. In: *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 1, Long papers, vol 1*, pp 1107–1116
7. Dauphin YN, Fan A, Auli M, Grangier D (2016) Language modeling with gated convolutional networks. [arXiv:1612.08083](https://arxiv.org/abs/1612.08083)
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
9. Grave E, Joulin A, Cissé M, Grangier D, Jégou H (2016) Efficient softmax approximation for GPUs. [arXiv:1609.04309](https://arxiv.org/abs/1609.04309)
10. Xiao Y, Cho K (2016) Efficient character-level document classification by combining convolution and recurrent layers. [arXiv:1602.00367](https://arxiv.org/abs/1602.00367)

11. Kim Y (2014) Convolutional neural networks for sentence classification. [arXiv:1408.5882](#)
12. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. [arXiv:1404.2188](#)
13. Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. *J Mach Learn Res* 3:1137–1155

Algorithm Research of Face Recognition System Based on Haar



Xiaoguang Deng, Zijiang Zhu, Jing Chang, and Xiaojing Ding

Abstract Face detection is to detect the face from the original image, which is the basis of face recognition technology. Face is a kind of natural structural object with quite complex details, which will be challenges to this kind of recognition. Face has the variability of pattern because of the different appearance, expression and skin color, and in general, there may be appendages such as glasses, beard and so on. Face image as a three-dimensional object has shadows caused by light also unavoidable. In this paper, the most mainstream Harr feature classifier algorithm is used, innovatively propose non-human rejection algorithm to reduce the impact of complex environment on detection results and improve detection accuracy.

Keywords Face recognition · Haar feature · PCA-LDA · Non-human rejection algorithm

1 Introduction

The research on automatic face detection and recognition originated in the 1960s. The representative achievement is the technical report published by Chan et al. on Panoramic Research Incorporated in 1965 [1]. In the past decades, great progress has been made both in technology research and technology application in the field of face recognition. In recent years, the research on linear discriminant analysis of face images has aroused widespread interest, focusing on how to extract effective discriminant features and reduce dimension. The task of feature extraction research is to find the most discriminant description of patterns and maximize distinguish from other types of patterns; to compress the dimension of schema data description

X. Deng (✉) · Z. Zhu · J. Chang
South China Business College, Guangdong University of Foreign Studies, Guangzhou, China
e-mail: 94143082@qq.com

X. Ding
China Baidu Ltd., Beijing, China

under appropriate situations, which is very important even indispensable when the original data space of the description pattern has a larger dimension.

Face recognition is an important topic in the field of pattern recognition, and it is also a very active research direction at present. It can be generally described as [2]: to determine whether there is one or more faces in any given image or video, to extract facial feature data and use the existing face database to identify one or more people from the image or video. Face detection is the basis of recognition in face recognition, how to detect face quickly and accurately in complex image or video environment is an important measure and premise to improve the efficiency of the system.

This paper mainly studies the optimization of non-human rejection algorithm when detect face features extracted by PCA-LDA algorithm, to reduce the false recognition rate of face detection and improve the recognition rate of face, so as to reduce the impact of external factors on the accuracy of face recognition.

2 Algorithmic Optimization

2.1 Haar Features

Haar feature is a simple rectangular feature proposed by Viola et al. [3]. It is named for its similarity to Haar wavelet, which was first applied to face representation by Papageorgiou et al. Haar features are classified into four categories [4]: edge features, linear features, central features and diagonal features. These features are extracted with fixed feature templates.

As shown in Fig. 1, the feature templates contain white and black rectangles. The Haar feature refers to the difference between the sum of gray levels of the corresponding regions in image sub-window of the black rectangle and the white rectangle [5].

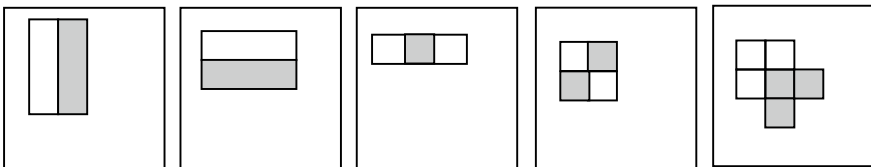


Fig. 1 Haar features

2.2 PCA Algorithm

The PCA algorithm [6] provides a linear transformation matrix between high and low dimensions, which can be obtained by computing the eigenvectors of the covariance matrix without other parameters. The reversible linear transformation matrix of PCA aims at minimizing the reconstruction error on the Euclidean distance of each dimension feature, and treats each dimension feature equally to intercept some part of the image energy, to reduce the dimension of the acquired data and improve the speed and accuracy of recognition.

Principal Component Analysis (PCA) is to try to combine many original related indicators (such as P) into a new set of independent comprehensive indicators to replace the original indicators. Usually, mathematical processing is to combine the original P index into linear combination as a new comprehensive index. The classic approach is to use the variance of F_1 (the first linear combination selected, that is, the first comprehensive index), which means that the larger the $\text{var}(F_1)$, the more information F_1 contains. Therefore, F_1 selected from all linear combinations should have the largest variance, so F_1 is called the first principal component. If the first principal component is not enough to represent the original P index information, then consider selecting F_2 , that is, selecting the second linear combination. In order to effectively reflect the original information, the existing information of F_1 does not need to appear in F_2 . To express in mathematical language is to require $\text{Cov}(F_1, F_2) = 0$, F_2 is called the second principal component. Analogous to construct the third, the fourth... the P principal component.

$$F_p = a_{1i} * z_{x1} + a_{2i} * z_{x2} + \dots + a_{pi} * z_{xp} \tag{1}$$

$a_{1i}, a_{2i}, \dots, a_{pi}$ ($i = 1, \dots, m$) is feature vector corresponding eigenvalues of covariance matrix Σ for X ; $z_{x1}, z_{x2}, \dots, z_{xp}$ is the value of original variable after standardizing disposed. Because of different dimensions of indicators in practical application, the effect of dimension must be eliminated first before calculation to standardize the original data, the data used in this paper has dimension effect [Note: data standardization in this paper refers to Z standardization].

$$A = (a_{ij}) * P * m = (a_1, a_2 \dots a_m), Rai = \lambda_i * a_i \tag{2}$$

R is a correlation coefficient matrix, λ_i, a_i is the corresponding eigenvalue and unit eigenvector, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

The main steps of principal component analysis are as follows:

1. Standardization of index data (automatic execution of SPSS software);
2. Determine the correlation between indicators;
3. Determine the number of principal components;
4. Principal component F_i expression;
5. Principal component F_i naming.

According to the feature extraction method of the above PCA algorithm, we can find that PCA algorithm is based on the gray statistics of face image. If the image is large affected by external factors or the face itself has great differences, then PCA algorithm cannot distinguish effectively. Therefore, external factors such as pose, expression, illumination, angle and size will greatly reduce the recognition rate of human faces. Secondly, the dimension of image vectors obtained by PCA algorithm is generally high, which result in high computational complexity in feature extraction. Finally, PCA algorithm cannot improve the recognition rate of images by training the class information of samples.

2.3 PCA-LDA Algorithm

Linear Discriminant Analysis (LDA), also known as Fisher Linear Discriminant (FLD) [7], aims to identify and extract the most discriminant low-dimensional features from high-dimensional feature space. These features can help to group all samples of the same class together and separate the samples of different classes as much as possible, that is, to select the feature that maximizes the ratio of inter-class and intra-class dispersion S_b . Intra-class dispersion matrix is to find the average value of each class of training samples, and then subtract the mean value of each class with each sample.

According to the characteristics of PCA algorithm and LDA algorithm, this paper uses a PCA and LDA fusion algorithm (PCA-LDA Algorithm) for feature extraction. The projection lines obtained by PCA method make the projected samples no longer separable, but the projection lines obtained by LDA method make the projected samples still have good separability. After projection, the samples of different classes in the low-dimensional space can be divided as far as possible, at the same time, we hope that the samples within each class are as dense as possible. That means the better the greater dispersion between samples, the better the smaller dispersion within the sample class. Therefore, if S is attached to a nonsingular matrix, the optimal projection direction W is the orthogonal eigenvectors maximize determinant ratio of discreteness matrix between sample classes and discreteness matrix within sample classes.

PCA-LDA algorithm process as follows:

- (1) Use PCA method to get a person's face space, which is the characteristic face space;
- (2) Calculate the characteristic subspace of LDA algorithm on this basis;
- (3) The generalized eigenvalues and eigenvectors are combined to form the optimal classification space;
- (4) Fusion of PCA algorithm and LDA algorithm feature subspace to get the fusion feature space of PCA LDA algorithm;

- (5) The training samples and the test samples are projected into the fused feature space to get the recognition features, and gender identification can be completed by using the nearest neighbor rules.

2.4 Non-human Rejection Algorithm

Non-human rejection algorithm, to detect face by comparing the similarity between the detected face and the standard face, eliminating the interference caused by mountains, light, buildings and so on, to solve the problem of a large number of false face detection. The principles are as follows:

Firstly, the Haar features of all faces in the face database are extracted, and the radius R_f of all faces is calculated. If a face is detected, then further calculate the radius of the face. If the radius of the face is less than or equal to the radius R_f of the face, then the recognized face is considered as a face.

Secondly, the Haar features of all non-faces in non-face database are extracted, and the radius R_{uf} of all non-faces is calculated. If a non-face is detected, then the non-face radius is further calculated. If the non-face radius is greater than or equal to the non-face radius R_{uf} , then the recognized non-face is considered to be non-face.

The calculation methods of single face radius and all face radius R_f are as follows:

1. Calculating the Radius of a Single Face
 - (1) Calculate the projection coordinates of a single face
 - (2) Calculate the Center of a single face
 - (3) Calculate the radius of a single face to the center, that is, the radius of a single face.
2. Calculating the Radius R_f of all Faces
 - (1) Calculate the projection coordinates of all faces
 - (2) Calculate the center of all faces
 - (3) Calculate the radius of the same person to all face centers
 - (4) Calculate the average radius of all faces, that is, R_f for all faces.

3 Experimental Results and Analysis

The facial database used in this experiment is a hybrid facial database, which contains 400 single-face images, 200 three-face images and 200 landscape images.

Among them, the ORL face database produced by AT&T Laboratory of Cambridge University in UK is used for single-face image sample. Each face image has 256 gray levels, and its original size is 112*92, as shown in the Fig. 2.

In this paper, we use 100 feature space dimension to recognize the single-face image and landscape image, multi-face image and landscape image respectively when



Fig. 2 Sample image of ORL face database

Table 1 Effect of non-human rejection algorithm on face recognition in single-face image

Non-human rejection algorithm	Misunderstanding number	Misunderstanding rate (%)	Recognition rate (%)
Close	4	1.00	94.25
Open	1	0.25	94.75

Table 2 Effect of non-human rejection algorithm on face recognition in multi-face images

Non-human rejection algorithm	Misunderstanding number	Misunderstanding rate (%)	Recognition rate (%)
Close	37	6.17	88.50
Open	7	1.17	90.67

non-human rejection algorithm is turned on or off in the system. The experimental results are shown in Tables 1 and 2.

From the experimental data in Tables 1 and 2, it can be concluded that:

- (1) In face recognition of 400 single-face images and 200 landscape images, non-human rejection algorithm has only a slight impact on the recognition results, reducing the number of false recognition from the original four to one, reducing the false recognition rate, but the number of recognized faces has not been greatly improved, and the recognition rate has not been greatly improved.
- (2) In face recognition of 200 three-face images and 200 landscape images, there are a lot of mistaken numbers in the recognized faces because of the complexity of face images. Although the non-human rejection algorithm does not significantly improve the system's face recognition rate, it greatly reduces the false recognition rate of the face, making only 7 faces recognized wrong from 551 images, the recognition rate of the system has improved.

4 Conclusion

The experimental results show that the proposed non-human rejection algorithm does not improve the system's face recognition rate dramatically, but it weakens the impact of environment and face numbers in the process of complex multi-face recognition, reduces the number of non-faces and the false recognition rate of faces effectively, the recognition accuracy of the system is improved.

Recognition Rate Can Be Greatly Improved

Acknowledgements This work was supported in part by a grant from the Youth Innovation Talent Project of colleges and universities of Guangdong Province China (Natural Science, No. 2016KQNCX230, 2016).

References

1. Bledsoe WW, Chan H (1965) A man-machine facial recognition system: some preliminary results. Technical report, PRI 19A. Panoramic Research Incorporated, Palo Alto, USA
2. Xu Q, Shi Y (2008) A face detection method based on feature model. *Comput Eng Appl* 44(14):202–204
3. Viola P, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the international conference on computer vision and pattern recognition*, pp 511–518. <https://doi.org/10.1109/cvpr.2001.990517>
4. Jiang Z, Chen C (2017) Face feature extraction algorithm based on Haar features and improved HOG. *Comput Sci* 44(1):303–307
5. Zhou SR, Yin JP (2013) LBP texture feature based on Haar characteristics. *Ruan Jian Xue Bao/J Softw* 24(8):1909–1926 (in Chinese). <http://www.jos.org.cn/1000-9825/4277.htm>
6. Yin F, Feng D (2008) Face recognition based on PCA algorithm. *Comput Technol Dev* 18(10):31–33
7. Lihamu YI, Ermaimaiti YA (2013) Face recognition based on improved PCA and LDA fusion algorithm. *Comput Simul* 30(1):415–418

Personal Authentication Based on EEG Signal and Deep Learning



Gi-Chul Yang

Abstract Personal authentication is an essential tool in this complex and modern digital information society. Traditionally, the most general mechanism of authentication was alphanumeric passwords. However, passwords those are hard to guess or to break, are often hard to remember. There are demands for a technology capable of replacing the text-based password system. This paper introduces a personal authentication system using a machine learning technique with EEG signal as a new type of personal authentication system which is easy for users to use and is difficult for others to steal.

Keywords Electroencephalography · Information security · Machine learning · Personal authentication

1 Introduction

According to Wikipedia “Authentication (from Greek: ἀθηντικός *authentikos*, “real, genuine”, from ἀθέντης *authentēs*, “author”) is the act of confirming the truth of an attribute of a single piece of data claimed true by an entity” [1]. Personal authentication is the process of confirming the identity of a person and an essential tool for everybody in this modern information society. The importance of the security of various digital devices increases day-by-day.

Traditionally, the most general mechanism of authentication was alphanumeric passwords. Passwords are a convenient and efficient scheme of authentication, but they do have drawbacks. Passwords should be easy to remember, but hard to break. However, passwords those are hard to guess or to break, are often hard to remember. That is one of the big problems of text-based passwords. An emerging research topic on personal authentication focuses on developing a secure and user-friendly authentication system.

G.-C. Yang (✉)

Department of Convergence Software, Mokpo National University, Mokpo, Korea
e-mail: gcyang@mokpo.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_45

325

Graphical passwords have begun to be used in authentication systems based on the fact that people can remember images better than text [1]. However, graphical passwords, which are generally easy to remember, have the disadvantage of being easy to steal from others. Therefore, this paper introduces a reliable password system using Brain-Computer Interface (BCI) as a new type of personal authentication system which is easy for users to use and is difficult for others to steal.

Since Hans Berger's discovery of brainwaves [2], EEG has long been used mainly in hospitals and laboratories to assess neurological disorders, to investigate brain function, and several studies have explored the possibility of treatment [3–5]. As the research progresses, it has developed into a research that can interpret the brain waves to read other people's thoughts, and use them to adjust peripheral devices or communicate with others [6]. Recently, researches on brain-computer interface have been actively conducted in various areas. This paper introduces a password system based on electroencephalography (EEG) and deep learning technique for personal authentication.

Using brainwave as a password has the advantage that it is hard for others to steal. However, it is difficult to guarantee its reliability because of the technical limitations of current brainwave signal interpretation. Therefore, in order to guarantee the reliability of brainwave information, this study suggests developing a system using EEG based deep learning technique. This idea makes it possible to develop a reliable password system that utilizes the difficulty of stealing brain information.

In the next section, we will learn more about EEG signals, and in Sect. 3 we introduce a machine learning-based password system. And conclude the paper in Sect. 4.

2 EEG Signals

Brainwave or EEG is a change in current due to the electrical current that flows when a signal is transmitted between neurons inside the brain. The wavelength of the EEG from the human brain is basically a frequency of 0–30 Hz and has amplitude of about 20–200 μ V. The frequency range of the EEG is arbitrarily classified into a delta wave with a frequency of less than 4 Hz, a theta wave between 4 and 7 Hz, an alpha wave between 8 and 13 Hz, a beta wave between 13 and 30 Hz, and above 30 Hz is called gamma wave.

Alpha waves appear mainly in the occipital region, and may extend to the parietal and posterior temporal lobes. The alpha waves of the occipital lobe increase in amplitude when you close your eyes. Beta waves occur primarily in the frontal lobe. Beta waves are mainly signaling in the frontal-central regions with signals related to concentration. Electromyogram (EMG) is similar in frequency domain to beta wave. Theta wave can appear in various areas of the brain. Theta wave occurring in the median frontal cortex occurs mainly during cognitive processing and increases with memory load. Theta-waves measured at electrodes other than Fz or Pz are likely to consider as abnormal waves. Gamma waves are thought to be related to higher

Table 1 The characteristics of EEGs

EEG	Frequency domain	Characteristics
Delta (δ) wave	0.2–3.99 Hz	– Occurs when sleeping
Theta (θ) wave	4–7.99 Hz	– Occurs when falling asleep
Alpha (α) wave	8–12.99 Hz	– Occurs when you are stable
SMR wave	12–15 Hz	– Occurs when attention is focused – When the efficiency of work is optimal
Beta (β) wave	13–30 Hz	– Occurs when mental activity or nervous system is active – Occurs during activities such as anxiety and tension
Gamma (γ) wave	30 Hz or higher	– Occurs in extreme arousal and excitement

cognitive functions such as the integration of cognitions. Gamma waves are closely related to concentration and memory, and are activated in the occipital lobe area when visual stimuli are meaningful (Table 1).

3 EEG Based Machine Learning

The EEG password system should be able to measure a user’s intention with a simple and inexpensive EEG device, so that a practical password system can be developed. Therefore, analysis and classification of the accepted EEG signals are important to distinguish certain user’s EEG signals among others. If we can distinguish each user’s EEG signal correctly, we can use it as a password. The system proposed in this paper is working based on EEG signals and called PassEEG.

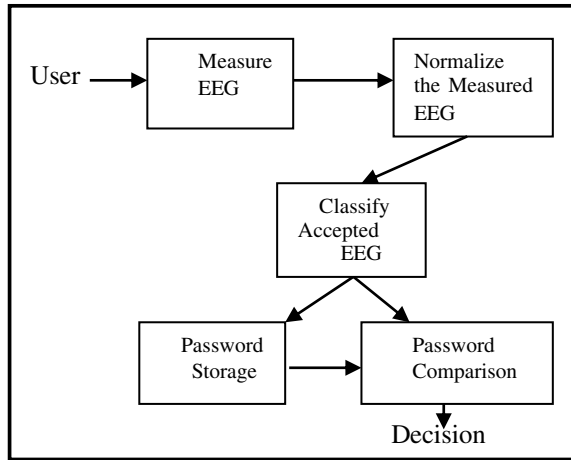
PassEEG accepts EEG signals from a user and normalize them to use it as personal information for authentication. After the normalization of the signal values, classify it by using multinomial classification. For example, the Hypothesis for multinomial classification can be:

$$H(x_1, x_2, x_3) = w_1x_1 + w_2x_2 + w_3x_3 + b$$

In the process of multinomial classification apply one-hot encoding technique in order to select final one category among many different categories. Next, store classified EEG signal values in the password storage along with the user ID to register a user’s password. All the process for login is same with the registration process but the last step. For login, the incoming classified EEG signal is compared with the signals stored in the password storage to identify a certain user. Figure 1 is a brief diagram of the proposed system.

The EEG-based password system is mainly divided into steps of generating and storing EEG and classification. To register the password, the user measures the EEG using the EEG device and classifies the measured signal according to a deep learning

Fig. 1 System configuration of the proposed system



technique and stores it in the password database with the user ID. For password authentication, EEG is measured using an EEG device and classify a input password to compared with a password stored with the same ID to determine whether authenticate it or not.

The important point here is whether you can recognize the password. The EEG password system should be able to measure a user's intention with a simple and inexpensive EEG device, so that a practical password system can be constructed. Therefore, in order to guarantee reliable EEG measurement, we propose to use an incremental password stacking method to accept newly classified EEG password without making any conflict with existing passwords. Among the deep learning technique, we recommend the multinomial classification with one-hot encoding technique in order to accurately recognize user's intention.

4 Conclusion

In this paper, we proposed an EEG-based password system using EEG and deep learning technique. EEG-based Password System makes it difficult for others to steal user's password, but it is unreliable to use in reality due to the limitation of current brain-computer interface technology. To overcome this problem, this paper proposed an EEG-based password system that uses not only EEG but also a deep learning technique. The proposed deep learning technique is multinomial classification with one-hot encoding. Therefore, using EEG and deep learning technique together is very efficient and useful to grasp user's intention. In the future research, we will implement the EEG-based system using the proposed method and show the test results.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) and funded by the Ministry of Education, Science

and Technology (2017R1D1A1B04032968). This research was studied by Research Funds of MNU Innovative programs for University in 2019.

References

1. Blonder G (1996) Graphical passwords. Lucent Technologies, Inc., Murray Hill, NJ, U. S. Patent, Ed. United States
2. Berger H (1929) Uber das electrenkephalo-gramm des menchen. Arch Psychiatr Nervenkr 87:527–570
3. Travis TA, Kondo CY, Knott JR (1975) Alpha enhancement research: a review. Biol Psychiatry 10:69–89
4. Rockstroh B, Elbert T, Canavan A, Lutzenberger W, Birbaumer N (1989) Slow cortical potentials and behavior, 2nd edn. Urban and Schwarzenberg, Baltimore, MD
5. Serman MB (2000) Basic concepts and clinical findings in the treatment of seizure disorders with EEG operant conditioning. Clin Neurophysiol 31:45–55
6. Hwang H-J et al (2013) Brain-computer interfaces: a thorough literature survey. Int J Human-Comput Interact 29:814–826
7. Hwang HJ, Lim JH, Jung YJ, Choi H, Lee SW, Im CH (2012) Development of an SSVEP-based BCI spelling system adopting a QWERTY-style LED keyboard. J Neurosci Methods 208:59–65
8. Volosyak I (2011) SSVEP-based Bremen-BCI interface-boosting information transfer rates. J Neural Eng 8:036020

Security Information and Event Management Model Based on Defense-in-Depth Strategy for Vital Digital Assets in Nuclear Facilities



Sangwoo Kim, Seung-min Kim, Ki-haeng Nam, Seonuk Kim,
and Kook-huei Kwon

Abstract After striking event of Stuxnet in Iran, international society recognizes that sabotage using cyber attack on nuclear facilities is no longer a hypothetical. The International Atomic Energy Agency, the IAEA, and the US nuclear facility regulatory authority recommends that nuclear licensees establish security measures to prevent/detect/respond the cyber attack. Moreover, storing logs at the system to trace and support the incident investigation and analysis by their guidelines. In particular, since vital digital assets (VDA) to prevent and mitigate severe accidents in nuclear facilities, that possibly be direct targets for sabotage. Therefore, security measures for cyber attack detection and log collection are essential. SIEM is typical attack detection model through security information and log management, and various solutions are already used in many IT industries. But VDAs are difficult to purchase and implement commercial log collection and detection solutions. Because industrial control systems which used in VDAs are develop specifically for nuclear facilities, designed and performing safety and safety related functions. And nuclear facilities are necessary to meet safety and security requirements such as defense-in-depth strategy and boundary protection system to licensees designing SIEN network to implement central monitoring method. So we proposed DID-SIEM that is a security information and event management model based on defense-in-depth strategy.

S. Kim (✉) · S. Kim · K. Nam · S. Kim · K. Kwon
Korea Institute of Nuclear Nonproliferation and Control, 1534, Youseong-daero, Youseong-Gu,
Daejeon, Korea
e-mail: kjoey@kinac.re.kr

S. Kim
e-mail: smkim90@kinac.re.kr

K. Nam
e-mail: kihaeng.nam@kinac.re.kr

S. Kim
e-mail: sw907@kinac.re.kr

K. Kwon
e-mail: vivacita@kinac.re.kr

DID-SIEM is SIEM model that incorporates the design requirements to meet both the cyber security guidelines and operational constraints of nuclear facilities.

Keywords SIEM · Defense-in-depth · Nuclear facilities · Security event · Critical infrastructure

1 Introduction

Recently, an unknown malicious code aimed at playing sabotage Iran's Stuxnet and cyber attack to KHNP, and thus interest for nuclear facilities in cyber security increased [1]. Vital Digital Assets (VDA) are critical systems for safety, so continuous monitoring and detection the abnormality of operating function is essential. This is because why the digital assets in nuclear facilities are prevented from malfunctioning of the control system and failure to mitigate accidents due to cyberattacks after the initial event [2, 3]. Regulatory guidelines in the United States and South Korea recommend collecting and analyzing logs to detect signs of cyberattack or anomalous behavior of systems caused by malware in order to prevent radiological consequence in nuclear facilities [4, 5]. The International Atomic Energy Agency (IAEA) cybersecurity guideline "Computer Security at Nuclear Facilities (NSS.17)" also recommends that the security-related logs are to be recorded [6]. Security Information and Event Management (SIEM) is a model that conducts cyber attack detection and manage the collecting and analyzing security information and events. Researchers are trying to conducted cyber security control and applied it to infrastructure [7–9]. SIEM centrally collects, analyzes, and monitors the network security information and security logs, and system information on each endpoint in the operation environment to provide an environment to timely detect cyber attack and abnormal symptoms. To build a system, such a SIEM, it is essential to configure network environment with a central server for collecting and managing each endpoint and security information. However, there are constraints on network configuration because control systems in nuclear facilities, including VDAs, must comply with physical and logical closure and cybersecurity requirements. This paper analyzed the specificity of nuclear facilities and the constraints that should be considered in applying SIEM, mainly by analyzing IAEA NSS.17, "NRC Regulatory Guideline (NRC RG 5.71)" and "Regulatory Standard-Security for Computer and Information System of Nuclear Facilities (KINAC RS-015)", the Korean cyber security regulation. Based on this, the requirements for the SIEM construction of nuclear facilities were identified. Finally, this paper proposed the architecture and design requirements of DID-SIEM that present as SIEM at nuclear facilities, a security log collection and management model based on the defense-in-depth protection strategies that are to be considered in constructing networks in nuclear facilities. This study derives design requirements based on security requirements which referenced from regulatory guidelines and constraints from operating environment that must be considered when constructing SIEM at nuclear facilities.

2 Related Research

This chapter describes the function and configuration of VDAs and basic SIEM, which are the main monitoring targets of DID-SIEM.

2.1 Vital Digital Assets

In IAEA nss.17, RG 5.71, defines critical digital assets related to safety, safety, security and emergency response in nuclear facilities as a mandatory digital asset. It requires that it be applied to cybersecurity measures. VDAs are digital assets that can cause core damage through mitigation failure if they are cyber attacked after the initial event of nuclear power. Vital digital assets are based on the Probabilistic Safety Assessment (PSA) model that identifies a set of assets related to severe accidents in nuclear facilities, and uses risk-informed information to prevent nuclear accidents caused by cyber attacks as shown in Fig. 1. The PSA is a comprehensive stability assessment that includes equipment failures, power failures, and external environmental factors in nuclear power plants. The PSA identifies all critical accident scenarios that may occur in a nuclear power plant and quantitatively calculates the core damage frequency of the target nuclear power.

In Korea, this methodology is applied nuclear power plants. Vital digital assets are important devices for preventing and mitigating serious accidents such as core meltdown, and require stronger cybersecurity measures [2, 3]. In addition, safety function diversity may be composed of non-safety devices that are important for safety functions and thus those asset also need to be considered when designing SIEM model.

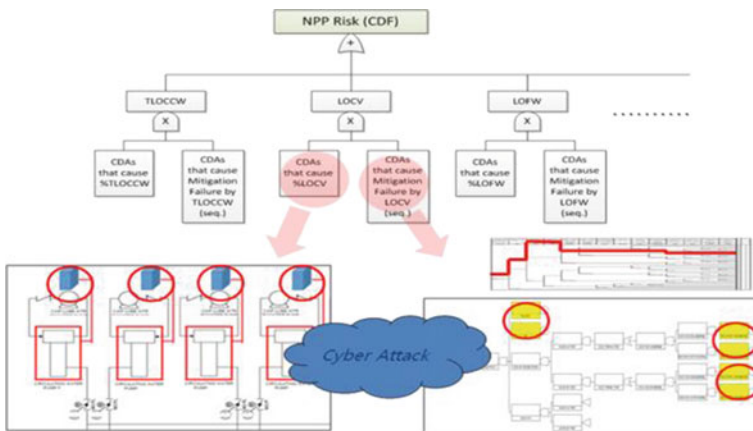


Fig. 1 Concept for identification method of VDAs

2.2 Security Information and Event Management (SIEM)

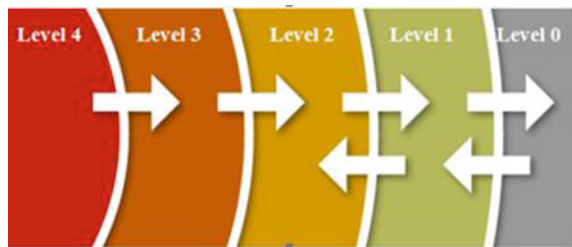
SIEM is a method recently used for security control by many large organizations and companies for identifying the immediate situation through centralized log collection and analysis. Security control through log collection consists of user agent, database, and collector to analyze and monitor depending on its function. The log collection server and database collect log information in real time according to a policy set from an agent installed in each host. The database stores log information in the database and provides the control system with the information needed for monitoring and analysis. The control system analyzes various event logs and provides analysis information such as attack detection result and traffic to CERT as such. Thus, the integrated comprehensive management system for collecting and analyzing security information from the network and security equipment is called SIEM [8, 10, 11].

2.3 Defense-in-Depth Strategy

Defense-in-depth strategy, the site security framework required by IAEA NSS.17 and RG 5.71 as the basis for building the SIEM network, is proposed in this paper. Nuclear facilities licensee shall apply the defense-in-depth strategy to protect the VDAs from the cyber attack including Design Basis Threats (DBT). Cyber security defense-in-depth structure is composed of cyber security defensive levels, and the security controls the cyber attack detection, prevention, delay, mitigation, and recovery.

Figure 2 shows an example of cyber security defense-in-depth structure. Such cyber security defense structure is composed of five cyber security levels divided by cyber security boundaries. The digital communication at each boundary is monitored and controlled. The higher cybersecurity asset should be located at the higher level in this model. The logical model suggested in (Fig. 2) is not directly related to the physical location [3, 6].

Fig. 2 Example of cyber security defense-in-depth structure



3 DID-SIEM

3.1 Constraints in Designing Centre Log Monitoring System for Nuclear Facilities

In today’s IT systems, SIEM is a popular security technic, and various commercial or open source products such as Arcsight, Elastic-Logstash are used. However, computer systems used in the operation of nuclear facilities are very different environment from other general IT systems, and there are many environmental and technical constraints for commercial SIEM products. Nuclear facilities can cause more direct damage to human life or nature in comparison to another infrastructure and common IT systems during cyber or safety incidents. Therefore, in order to change the control system of nuclear facilities, it is necessary to perform safety verification such as software V&V (Validation and Verification) to ensure the safety function is maintained [12]. Because the safety verification process is costly and time consuming, most nuclear facility control systems use old computer system and industrial protocols designed for safety. In addition, in accordance with IAEA NSS.17, the network of CDA should be protected by applying access control. This requirement also applies to SIEMs that require network connectivity with CDA for log collection.

Regulatory guidelines in the IAEA, Korea and the U.S.A require nuclear licensees to implement a variety of technical security measures, including defense-in-depth, for cybersecurity of control systems.

Table 1 describes the constraints from operating environment and security requirements that must be considered in establishing a cybersecurity control using SIEM in the control system of nuclear facilities.

Table 1 Constraints of SIEM when it design for nuclear facilities

	Details
Constraints from operating environment	Various type of industrial protocols and platforms
	Only Authorized personal can access
	Insufficient H/W resources due to aged systems
Technical constraints based on security requirements	Network isolation between different level for defense-in-depth
	Safety function is the highest priority over security
	In case of access to the boundary protection system, use equipment designated in advance to perform direct access (remote access is prohibited)
	Access control to database that manage key
	Security function should be separated from other functions
Ref. IAEA NSS.17, NRC/RG 5.71, KINAC/RS-015	Apply other network security function (confidentiality, integrity, etc.)

3.2 *Design Requirements and Architecture of DID-SIEM*

This section describes the design requirements and architecture of the DID-SIEM to address the constraints in Table 1. The design requirements of DID-SIEM are classified into network, security management/monitoring, collection/analysis and host agent according to their function.

The design requirements for network construction of DID-SIEM are as follows:

- a. Network isolation (Safety, Non-Safety, Security Control)
- b. ICS network of nuclear facilities need to be isolate from office network and internet to remove external attack vectors (Air-gapped)
- c. Prohibit the data transfer from low security level to high security level
- d. Dividing security policy control and key distribute function from monitoring system
- e. One-way transmission through boundary protection system for data transmission from safety network to non-safety network.

DID-SIEM is based on defense-in-depth, a network security strategy for nuclear facilities. Safety and non-safety networks are required to be separated according to the DID strategy. This separation the network as requirement (a). and restricts data transfer from high to low by a one-way deterministic device. In addition, the control network was separated to designate the terminals accessible to the boundary protection system and the key distributor according to RG 5.71 and RS-015. When connecting to external network for cloud function or central monitoring platform, any connection point to other network such as office net or internet should not be exist.

The design requirements for the monitoring system and security policy control service as follows:

- a. Separation of control terminal and security policy/key to separate control terminal to satisfy function separation and in-depth protection according to authority
- b. The key used for encryption/authentication is transmitted from the management console to the safety log collector and from the safety log collector to the non-safety log collector.
- c. The security protocol used for sending security information uses the protocol available in One-way.
- d. Boundary protection system is designed to control only security management terminal.

RG 5.71 and RS-015 require to protect the database that stores the encryption key and restriction and management of the terminals which can access the boundary protection system. Since the monitoring device must receive both safety and unsafe data, it must be connected to the unsafe network, and the security management terminal must transmit security information such as security policy and encryption key to the safe/unsafe to protect it from the safety network or higher. So DID-SIEM separates security management terminal and monitoring system. The data flow of security information is as follows (Fig. 3).

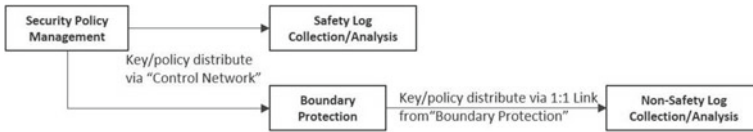


Fig. 3 Security key and policy data flow diagram

The design requirements for the collection and analysis system are as follows:

- a. Security information transfer from safety and non-safety peers is sent to the log collect system in each network.
- b. Log collection system existing in each network analyzes and processes the collected information to create meaningful security information.
- c. Selected and processed safety and non-safety security information and events are transfer to the monitoring system via the non-safety network.

First, each collection/analysis server is installed in each safety and non-safety network. Servers collect and analyze information from peers connected to each network. The requirement (a) prevents adversely affecting different levels of log collection if either system fails. Selected and processed safety and non-safety information is transferred to the monitoring system and the data flow is shown in Fig. 4.

The design requirements of the host agent are as follows:

- a. Agent for collecting security information of each host (Peer) should not generate its own log.
- b. Agent collects, processes and transmits logs provided by the platform.
- c. Prohibit any process that may affect the original function of the asset such as API hooking.
- d. Field devices using serial networks are converted to TCP/IP through HMI or Protocol Convertor.

The primary purpose of the control system, the host where logs are collected, is to perform safety and safety-related functions. Therefore, separate log collection and hooking that can affect system processing speed and process should be avoided. It

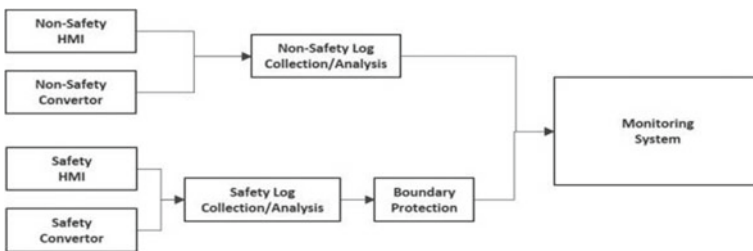


Fig. 4 Log data flow diagram

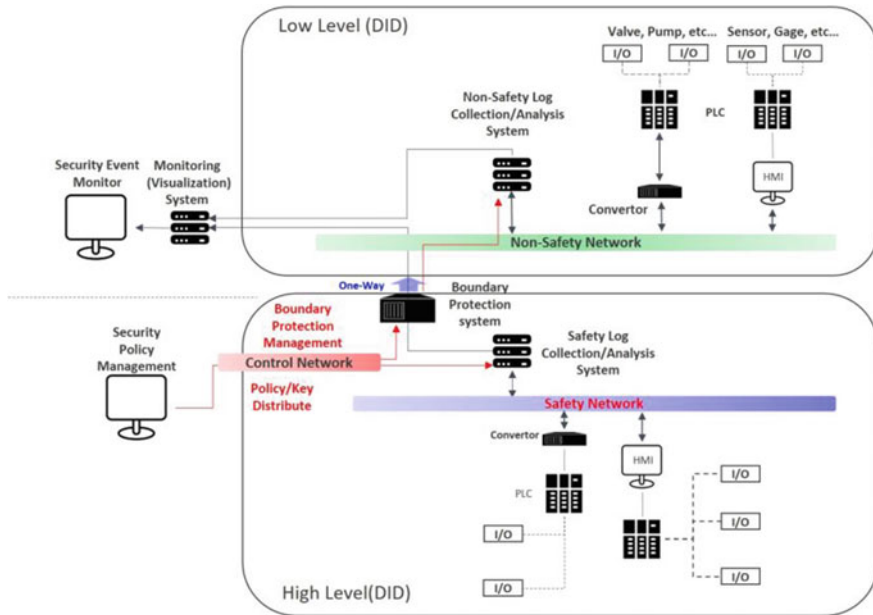


Fig. 5 DID-SIEM architecture

is necessary to unify the protocol to apply security measures such as confidentiality and integrity in log transmission. Requirement (d) is necessary to prevent the effects of SIEM network or system failures from propagating directly to the control system.

Figure 5 shows the suggested architecture of DID-SIEM reflecting the design requirements presented above.

4 Conclusion

Log management and control is necessary for cyber threat detection in nuclear facilities. However, it is difficult to apply the general SIEM model to nuclear facilities because it is difficult to meet the operational and technical security requirements presented in nuclear security guidelines. This study derives design requirements based on security requirements which referenced from regulatory guidelines and constraints from operating environment that must be considered when constructing SIEM for nuclear facilities. And proposes SIEM architecture applying the design requirements. DID-SIEM uses a network based on the DID strategy with network separation and data transmission restriction between different security levels. In addition, it is a model that reflects design requirements for designating terminals capable of boundary protection and security management and access control follows DID. This study is a basic technology research for event detection of vital digital assets,

and was conducted to help researchers who are considering SIEM for nuclear facilities. In the future, based on the opinions of nuclear operators and experts, the model will be improved, and safety and security evaluation will be conducted.

Acknowledgements This work was supported by the Nuclear Safety Research Program through the Korea Foundation Of Nuclear Safety (KoFONS), granted financial resource from the Nuclear Safety and Security Commission (NSSC), Republic of Korea (No. 1605007).

References

1. Lee S, Huh J-H (2018) An effective security measures for nuclear power plant using big data analysis approach. *J Supercomput* 1–28
2. Hwang M, Kwon K (2018) Development of an identification method for vital digital assets selection on nuclear cyber security. In: Transactions of the Korean nuclear society spring meeting
3. US Nuclear Regulatory Commission (2010) Regulatory guide 5.71. Cyber Security Programs for Nuclear Facilities, Washington, DC
4. KINAC, KINAC (2014) RS-015. Technical standard on cyber security for computer and information system of nuclear facilities
5. Kuipers D, Fabro M (2006) Control systems cyber security: defense in depth strategies. No. INL/EXT-06-11478. Idaho National Laboratory (INL)
6. IAEA, NSS (2011) No. 17: 2011. Computer security at nuclear facilities: reference manual: technical guidance. International Atomic Energy Agency, Vienna
7. Miller D (2011) Security information and event management (SIEM) implementation. McGraw-Hill
8. Coppolino L et al (2013) Enhancing SIEM technology to protect critical infrastructures. *Critical information infrastructures security*. Springer, Berlin, Heidelberg, pp 10–21
9. Novikova E, Kotenko I (2013) Analytical visualization techniques for security information and event management. In: 2013 21st Euromicro international conference on parallel, distributed, and network-based processing. IEEE
10. Bhatt S, Manadhata PK, Zomlot L (2014) The operational role of security information and event management systems. *IEEE Secur Priv* 12(5):35–41
11. Anastasov I, Davcev D (2014) SIEM implementation for global and distributed environments. In: 2014 world congress on computer applications and information systems (WCCAIS). IEEE
12. Gu P et al (2016) A study about safety I&C system software V&V in nuclear power plant. In: 2016 24th international conference on nuclear engineering. American Society of Mechanical Engineers

A Web Archiving Method for Preserving Content Integrity by Using Blockchain



Hyun Cheon Hwang, Ji Su Park, Byung Rae Lee, and Jin Gon Shon

Abstract A web archive system has become an essential topic for preserving historical information for descendants with the explosive growth of web data. The reference model for an Open Archival Information System (OAIS) has been providing an excellent guide for a long-term archiving system, and most of web archive systems follow this guide. However, there is still a weak point in terms of content integrity due to the archival web data could be altered by unauthorized manner. In this paper, we proposed the BCLinked (Blockchain Linked) web archiving method which uses blockchain technology and an extended WARC (Web ARChive) file format to ensure the content integrity. Furthermore, we confirmed the proposed method ensures content integrity through the experiment.

Keywords Web archive · OAIS · WARC · Blockchain · BCLinked web archiving method

1 Introduction

The purpose of the web archive is to collect web content and preserve it for the long term to deliver valuable data to descendants [1]. Consultative Committee for Space Data Systems (CCSDS) has developed Reference Model for an Open Archival

H. C. Hwang · B. R. Lee · J. G. Shon (✉)
Department of Computer Science, Graduate School, Korea National Open University, Seoul, Korea
e-mail: jgshon@knou.ac.kr

H. C. Hwang
e-mail: panty74@knou.ac.kr

B. R. Lee
e-mail: brlee@knou.ac.kr

J. S. Park
Convergence Institute, Dongguk University, Seoul, Korea
e-mail: bluejisu@dgu.edu

Information System (OAIS) which is ISO 14721 for long-term preservation of digital records, and many web archive systems follow the standard [2]. However, the OAIS reference model does not provide an implementation method but prescribe requirement to ensure OAIS-compliant [3]. Even though the OAIS reference model mentions the long-term preservation of digital records, the weakness of content integrity exists in current web archive system due to web retrieval content could be altered by unauthorized manner. Blockchain is the technology which connects a previous dataset by cryptography key and the whole dataset will be broken in case any dataset is changed [4]. Blockchain technology is getting widely used for IT industry who needs to keep all transaction records without harming. In this paper, we proposed the web archiving method based on blockchain technology which can use with WARC (Web ARChive) file format to enhance web content integrity.

2 Related Research

2.1 OAIS

OAIS Reference Model is technical guidance for a long-term archiving system. It developed by CCSDS in 2002 [3]. It is ISO 14721 Standard, and many digital archive systems follow this guide for their archive system infrastructure. The environment Model of an OAIS consists of Producers, Consumers, and Management [2]. Producers collect the data to be preserved and deliver to the OAIS as SIP (Submission Information Package). OAIS manages the data as AIP (Archival Information Package), and Consumers can use the archival data from OAIS. In the detailed description of functional entities of OAIS, there are three services guidance; Common Services, Network Services, and Security Services. Security Services tells data integrity service should ensure the data is not altered or destroyed in an unauthorized manner, and non-repudiation service should ensure to provide proof of the origin of data [2]. Even though there is strict security guidance security in OAIS, the reality is that the web archive system can't keep it all.

2.2 Web Archive

The purpose of web archive is the long-term preservation of current information from the web for descendants [1]. IIPC described four steps of the web archiving chain as shown in Fig. 1 [5]. The web archive is one of challenging topic due to the web data is increasing very fast, so various web archive collecting strategy has been developing by using web crawlers for the capture and archiving stage. IIPC has discussed WARC file format is standard ISO 28500 and it provides a standard way to structure, manage and store billions of resources collected from the web and

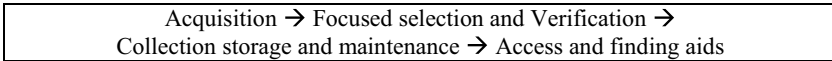


Fig. 1 Four steps of web archiving chain

elsewhere [6]. WARC file format is accepted in most of the web archive such as the national library of Korea [7].

WARC file format consists of more than one WARC records. In general, the first WARC record has the description for other records, and the records after second have the result of retrieval from a source such as a web page, images. There are the defined-fields in WARC record format to describe each record’s properties. There are two defined-fields for content integrity, which are WARC-Block-Digest and WARC-Payload-Digest [6]. However, these two fields are optional defined-fields, and anyone can change the digest value and the content together. It means there is no way to prove that the digest value and record value aren’t revised. So, the WARC file format has a weakness against unauthorized content modification.

2.3 Blockchain

Blockchain technology was developed by Satoshi Nakamoto for bitcoin in 2008 [4]. It uses dataset which called a block, and these blocks are linked with the previous block by using cryptography as shown in Fig. 2 [4]. This linked block set is called Blockchain. Each block contains the cryptography hash value of the previous block, a timestamp and a transaction dataset. Blockchain dataset stores across a peer to peer network to avoid centralization of holding dataset so that anyone can access the blockchain dataset to proof each block integrity. Because there is the possibility that any user in peer to peer network can create a new block and it can be a conflict with another user, Blockchain has the mechanism “proof of work”. The Blockchain is widely used for cryptocurrency, customer contract storing in FSI and etc. due to the characteristics of Blockchain such as (1) whole Blockchain will be broken in case

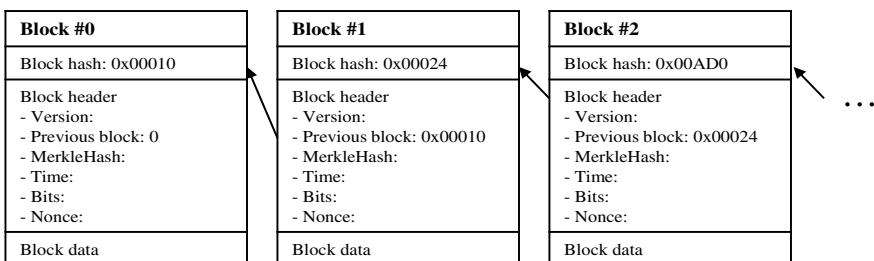


Fig. 2 Blockchain with linked blocks

anyone block is damaged, and (2) anyone who in peer to peer network can review the Blockchain dataset [8].

3 BCLinked Web Archiving Method

3.1 Web Archiving Method Based on Blockchain

One of the main problems in the current web archive system is that there is no central managed archive system as well as there is no content integrity proof system. Even though many organizations try to run a global standard web archive system for descendants, they run their system separately and it is very hard to synchronize to avoid duplicated web resources collection. Also, the web archive system needs to care of web archive transaction for each domain, however, there is no relationship among each domain. It should be enough that the web archive system can recognize the life cycle of each domain. So, we define the BCLinked (Blockchain Linked) web archiving method which has two levels blockchain for web archive as shown in Fig. 3. The blockchain node in the first level which named Domain Blockchain contains a domain name and new node will be added in case a new domain is found for web archive. The Domain Blockchain can be used for a web domain exists in web. The blockchain node in the second level which named WebContent Blockchain contains the linked information to the block node in the Domain Blockchain and the web archive retrieval information. The number of WebContent Blockchain will be the same as the number of nodes at the first level.

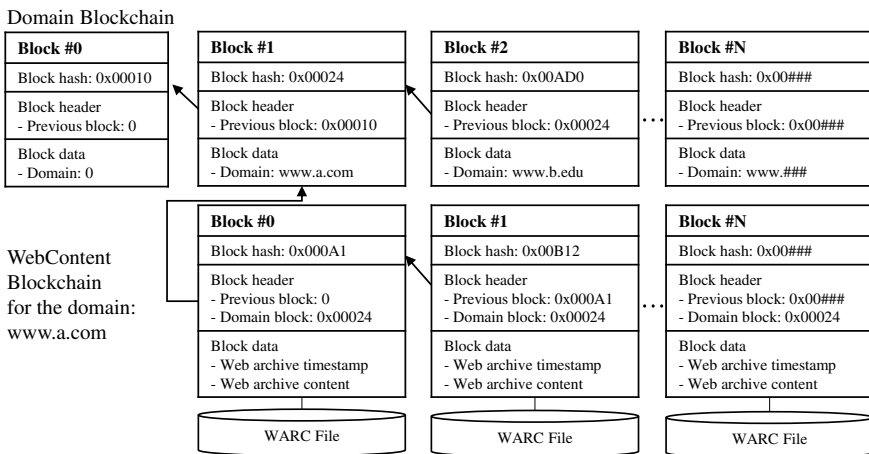


Fig. 3 Blockchain architecture used in the BCLinked web archiving method

Table 1 Definition of the Domain Blockchain and WebContent Blockchain node data

	Field	Value
Domain Blockchain	WARC-Domain	Web site domain name for archive
	WARC-Domain-CreationTime	Creation time of the node
WebContent Blockchain	WARC-Block-Digest	Block digest—algorithm “:” digest-value
	WARC-Payload-Digest	Payload digest—algorithm “:” digest-value

Table 2 Extended defined-field for WARC record

Field	Value
WARC-Added-Chain	True/False
WARC-Domain-Block	Block hash of the Domain blockchain
WARC-WebContent-Block	Block hash of the WebContent blockchain

The web retrieval result from web collection tools will be added to the blockchain of BCLinked web archiving method to ensure that the results are generated and the integrity of the results.

3.2 Extension of WARC Record Format

We define fields for blockchain node and WARC record to integrate with blockchain and WARC. The Domain Blockchain node will have each domain description and the WebContent Blockchain node will have each web retrieval result for each domain. So, we define the blockchain node data as shown in Table 1. All WARC record will be added to the WebContent Blockchain node separately. Only Block-digest and Payload-digest value will be added as blockchain node data due to all web retrieval result are too big for blockchain node data. Once adding digest value to the blockchain node, the description of the node will be added into a WARC record. We define the extended defined-fields for WARC record to describe the blockchain node as shown in Table 2.

3.3 Process of BCLinked Web Archiving Method

The process of BCLinked web archiving method is shown as Fig. 4. A web crawler collects web content and creates a WARC file in the content crawling stage. All

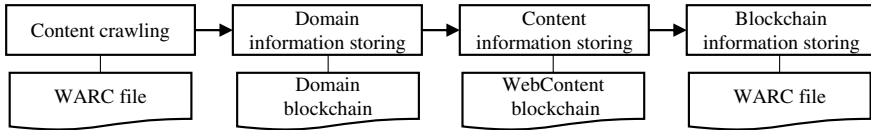


Fig. 4 Web data archival process in the BCLinked web archiving method

WARC records are sent to BCLinked web archiving. BCLinked web archiving method checks whether the target domain exists in the Domain Blockchain with the WARC records. The existing blockchain node data will be used if the domain already exists, or a new blockchain node will be added in the domain information storing stage. Then all WARC records are added to the WebContent Blockchain node in the content information storing stage, and the block hash is added into WARC records for cross-validation in the blockchain information storing stage.

4 Experiment and Analysis

We verify the BCLinked web archiving method described in this paper through this Experiment. We used the wget [9] for a web crawler and experimented the BCLinked web archiving method by python. We modified the archived web content to simulate how BCLinked web archiving method secure content integrity against unauthorized modification. The result summary is shown in Table 3. We confirmed the processing time of blockchain node calculation is a tiny portion of the whole process. About 98.4% time is consumed in the web crawling stage. The total storage consumption is around 0.2% higher than normal WARC archive system. In the next step, we selected one WARC file and modified the content. It took less than 0.5 s for a WARC file modification. But it took around 99.2 s for rebuilding the blockchain node data in BCLinked web archiving method in case the number of WARC is 100. The time rebuilding for blockchain node data can be increased linearly with the number of WARC files, and it is impossible to rebuild whole blockchain node data theoretically because the blockchain node data is stored in peer to peer network and the difficulty of calculation of blockchain can be adjusted.

Table 3 Comparison between a conventional web archiving and BCLinked web archiving methods

Archiving method	Conventional web archiving	BCLinked web archiving
Archive iteration	100 times (1 min per each)	100 times (1 min per each)
Processing time (s)	6,090	6,189
Total file size (byte)	43,877,480	43,994,500
Modification time (s)	0.5	99.2

5 Conclusion

In this paper, we proposed BCLinked web archiving method in order to preserving content integrity by using blockchain technology and extending WARC specification. According to the method, the archived web content is stored in WARC file and the block-digest of the web content is added to its blockchain node. With this block-digest value in blockchain node, we can verify the web content integrity preservation. In conclusion, BCLinked web archiving method is an enhanced web content collecting method by resolving the weakness in terms of content integrity. Especially, the proposed method can be much more valuable in case of a web archive system collecting integrity-sensitive content such as legally related content. In future research, we will try to find out how to get a specific version of web content efficiently.

References

1. www.wikipedia.org. Web archiving. https://en.wikipedia.org/wiki/Web_archiving. Accessed 14 Aug 2019
2. CCSDS. Consultative committee for space data system: reference model for an open archival information system (OAIS). <https://public.ccsds.org/pubs/650x0m2.pdf>. Accessed 14 Aug 2019
3. Park B-J, Cha S-J, Lee K-C. Data mapping between Korea deep web archiving format and reference model for OAIS, pp 197–200
4. Lemieux VL (2016) Trusting records: is blockchain technology the answer? *Rec Manag J* 26(2):110–139
5. Kim H-J (2011) Comparative analysis of web archiving tools. In: Korean Society of archives and records management conference, pp 95–98
6. <https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/>
7. www.oasis.go.kr. Website building guide”. <http://www.oasis.go.kr/about/guide.do>. Accessed 18 Aug 2019
8. Deloitte.com. Breaking blockchain open—2018 global blockchain survey. <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-2018-global-blockchain-survey-report.pdf>. Accessed 20 Aug 2019
9. www.wikipedia.org.wget. <https://en.wikipedia.org/wiki/Wget>. Accessed 18 Aug 2019

An SDN-Based Distributed Identifier Locator Separation Scheme for IoT Networks



Chan-Haeng Lee, Ji Su Park, and Jin Gon Shon

Abstract In the Internet of Things (IoT) environments, sensor devices need to be connected to the networks for data transmission and inter-communication with machine to machine, machine to human. Most devices connected to the network need a distinguishable identifier and a network address, therefore, IP address is the best solution for that. However, due to the constraints of sensor devices, it is hard to deploy on legacy IP system directly. To solve this issue, a concept for the separation of identifier and locator from the IP address is able to be the alternative. From the viewpoint of the addressing separation, we propose an SDN-based identifier locator separation architecture for IoT networks with distributed manner.

Keywords SDN · Software-defined network · IoT network · ID-LOC separating

1 Introduction

Recently, research on the Internet of Things (IoT), one of the major technologies of the 4th industrial revolution, has been actively conducted. The IoT technology is used to process meaningful or useful information in various industries using sensor devices and Information and Communication Technologies (ICTs). The IoT devices have to be connected to a network or Internet to provide information and connectivity. In the IoT environment, most of the sensor devices use an identifier (ID) rather than an Internet Protocol (IP) address for data transmission because of their constraints

C.-H. Lee

Division of General Studies, College of Liberal Arts and Interdisciplinary Studies, Kyonggi University, Suwon-si, Korea

J. S. Park

Convergence Institute, Dongguk University, Seoul, Korea

e-mail: bluejsu@dgu.edu

J. G. Shon (✉)

Department of Computer Science, Graduate School, Korea National Open University, Seoul, Korea

e-mail: jgshon@knou.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_48

such as power consumption, bandwidth limitation, and transmission rate per packet. The ID is used to identify each sensor device and to communicate with other devices. Data collected from the sensor can be delivered through a gateway connected to the network, and an IP address can be used to connect to an external network.

Most devices to connect to the Internet use IP address, and the IP address can be used as an ID and a locator (LOC) to bind a device and an application. However, the binding brings several disadvantages such as mobility, multihoming, and extensibility problems. Especially, the explosive increase of mobile devices causes their deployment and addressing problems on the network, nowadays. To provide mobility, scalable routing, and addressing, the ID and LOC of IP address need to be separated.

Several ID-LOC separation protocols are proposed [1–4]. These proposed protocols use two namespaces as ID and LOC, and they are classified into two types such as host-based protocols and network-based protocols. Host-based protocols, such as Host Identity Protocol (HIP) and Shim6, require the modification of protocol stack within the hosts and rendezvous server for maintaining ID and LOC mapping information, and they have initial deployment difficulty. In contrast, network-based schemes do not modify hosts' protocol stack [3, 4]. The mobile nodes (MNs) do not participate in any signaling procedure. The ID-LOC mapping is processed by other network components such as routers, and they forward packets using tunneling according to the mapping information.

However, they have problems such as bandwidth waste and processing overhead. Also, the host-based and network-based protocols require a centralized ID-LOC mapping system such as Rendezvous server, MAP server, and etc.

In this paper, we propose a Software-Defined Networking (SDN)-based ID-LOC separation scheme for IoT networks. To manage the IDs and LOCs, we adopt an overlaid network concept from content addressable network (CAN) in SDN [5–7]. The IDs are used to identify who the sensors or end-point nodes are, and the LOCs are used to identify where they are located in. In the proposed scheme, we suggest a field replacement function in the OpenFlow-enabled switches (OFSs) to reduce tunneling overheads.

This paper is organized as follows. Section 2 presents the proposed SDN-based ID-LOC separation scheme. In Sect. 3, we illustrate operation process of the proposed scheme. Section 4 provides conclusion and future works.

2 SDN-Based ID-LOC Separation Scheme

2.1 System Architecture

In the proposed scheme, it is assumed that a sensor node (SN) can send and receive data using IPv6 over low power wireless personal area network (6LoWPAN) protocol [8]. For the identifier and locator of a sensor node, EID and LOC are used respectively.

The EID is generated by address autoconfiguration specified in [8]. The EID is a stateless address which is based on the EUI-64 assigned to the IEEE 802.15.4 [8, 9].

The LOC is a routable IPv6 address and is generated by using current network prefix with a SN's layer 2 address based on the address auto configuration mechanism [10]. This LOC is used as a general IP address in an IP-based network and represents the current location of which a SN attached to.

In the proposed scheme, the network is partitioned into the Local and IP domain, and the EID domain is overlaid on the local domain. IPv6 is used for IP domains considering the scalability and compatibility for the legacy networks. The IP domain is operated for normal IP routing. The packet with a general IPv6 address will be forwarded using IP routing as SDN environments. The local domain is consisted of SNs, OpenFlow-enabled Switches (OFSs) and access routers (ARs) with an SDN controller function. SNs in the local domain transmits data via OFSs to the AR using wireless or wired connection. Each SN communicates using its EID which indicates each sensor separately. In the local domain, EID information is used to deliver sensor packets. Figure 1 shows the proposed network architecture.

In the EID domain, the virtual coordinate space is used to store and maintain the <EID, LOC> pairs on the controller in a local domain. The entire virtual coordinate space is dynamically divided into the number of controllers in the domain. A controller with EID management module is located in each zone, and is called Local Mapping Controller (LMC). Each LMC is responsible to manage the allocated EIDs and SNs which have to register EID-LOC mapping information.

All LMCs own their individual, distinct zone within the space, and they manage the <EID, LOC> pairs using DHT mechanism. The basic functions of LMCs are

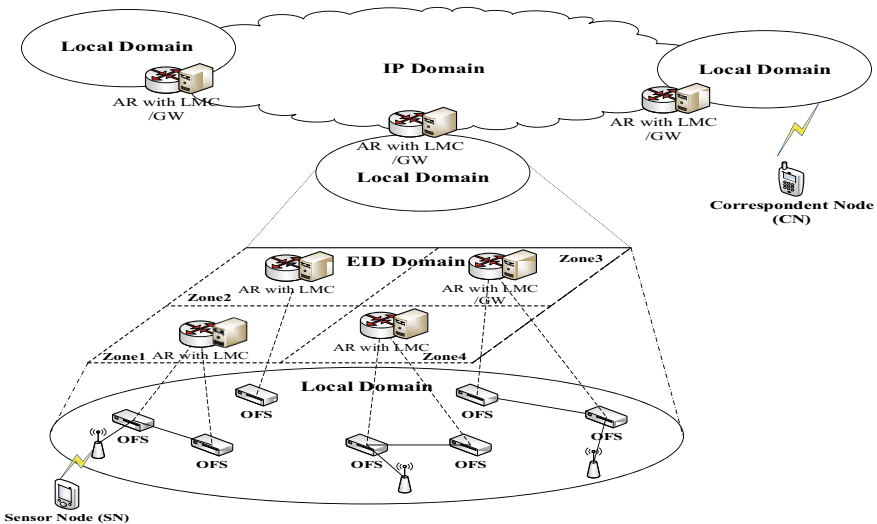


Fig. 1 System architecture of the proposed scheme

network attachment detection, mapping LOC to EID of nodes and updating the EID-LOC mapping, mapping information maintenance and management, and the flow table configuration of each OFS.

The ARs has three states according to their roles. The first is a home AR (h_AR) for a SN. If a SN's EID is included in an AR's EID block, this AR becomes a SN's h_AR . The h_AR has to manage the EID-LOC mapping information, and has the responsibility to send the corresponding LOCs for the EID queries. Second is the serving AR (s_AR) that manages a zone in which a SN is physically located. When the AR detects a new attachment due to the movement of the SN, it assigns a LOC for the EID of this SN and maintains the mapping information. It also announces and updates this EID-LOC binding information to the SN's h_AR .

When the SN having EID for its address tries to send a packet to a CN, the s_AR sends a query to get the LOC of CN. Once the s_AR obtains LOC information of the SN and CN, it sets the flow tables of its OFSs with the obtained information. The third is to act as a neighbor AR (n_AR). All the ARs except the h_AR and s_AR in a local domain are regarded as n_AR s. The n_AR s forward received packets to the nearest zone according to the preconfigured flow table for EID routing.

3 Operation Process

3.1 Registration

If a new SN tries to connect to the network, it has to register itself to the LMC. When the SN attached to the network, the nearest OFS detects the attachment. Generally, OFSs forward the received packets based on the matching rules of its flow table. If there is no matching from the received packet, the OFS sends this packet to its AR. Otherwise, it performs proper process according to the rule. The OFS has no information for the SN, it notifies the attachment to its AR, and then the AR also detects L2 attachment. From the L2 attachment, the AR can acquire the SN's ID and it generates the EID and LOC of the SN. The AR add a mapping table for the EID-LOC mapping information to its LMC. And it sends a command to add a flow table entry to its OFSs. This flow table entry is used for the ingressive LOC packets to exchange LOC for SN's EID.

In the proposed scheme, we add a function that converts LOCs to EIDs for the destination address to the SDN components. The address field conversion makes possible to find the optimal paths over IP domain, and forwards the packet through the paths. Therefore, the end nodes such as sensors, CNs can use their EID for the packet transmission.

3.2 Packet Forwarding

When the SN having EID for its address tries to send a packet to a Correspond Node (CN) who has an EID address, it first needs to figure out where the CN located. If the OFS which received this packet has no entries for this CN, it sends a packet-in message to its AR. The AR received the packet-in message also looks up its LMC to find the table entry first. If the AR does not have any entries for the LOC, then it sends a query to the other LMCs in the control domain for getting the CN's LOC. The AR also sends this binding information to its GW having LMC to indicate where the packets to go. This query message is forwarded using dedicated channel to the ARs. Once the AR of the SN obtains LOC information of the SN and CN, it sets an EID-LOC mapping table in the LMC and sends a command to update the flow table of its OFSs with the obtained information.

The OFS in the CN's AR already has the flow table entry to the CN by the registration process, the OFS performs the action in the flow table that replace the EID_CN in destination address field to LOC_CN. After the field replacement, the OFS send the packet to the output port of IP domain for normal IP routing.

The packet will be arrived at the CN's AR via IP domain, and the field replacement is occurred again from LOC_CN to EID_CN at the OFS of CN's AR. After that, the packet is forwarded to the CN.

The routes for the packet from SN to CN are established after the first packet arrived, and subsequent packets will be followed the configured routes.

3.3 Route Optimization

When the afterward packets are forwarded using the established routes, the packets are always passed to the attached OFSs of a SN and a CN's ARs through the routes which include several OFSs and ARs in the local domain. The routes forwarded via IP domain are optimized by using the IP routing, but the whole routes from a SN to a CN may not be optimized. If the OFSs which connected to the SN and CN have a suitable flow table entry for forwarding packets via IP domain, it is obvious that the whole paths between the SN and the CN will be optimized.

For providing optimized routes between a SN and a CN, a new rout optimization (RO) operation is required. At the packet forwarding process, the s_AR does not have information about CN, it sets a flow table entry to its OFSs and forwards the packet to the n_AR which is the nearest AR toward to the destination. The s_AR also sends a query request to the CN for obtaining the CN's LOC.

If the CN is in the local domain, the n_ARs forward the query request to the CN's AR, similar with the registration process. After the CN's AR receives the query request, it sends a reply with LOC information to the s_AR. When the s_AR receives the reply, it can save the SN's <EID, LOC> mapping information. It also sends a command to the OFS which connected to the SN for adding or modifying the

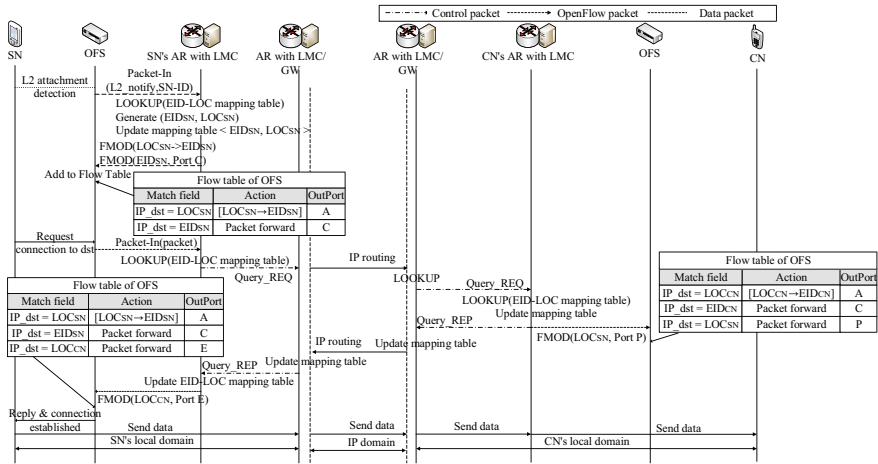


Fig. 2 Registration and packet forwarding process in the proposed scheme

flow table entries. The command includes the action set which converts destination address field from CN's EID to CN's LOC and changes the out port of packets.

If the CN is not a member of the local domain, that is, the CN is located in the other domain, s_AR requests query for the CN to the other domain through the IP domain. In this case, the CN's EID cannot be used to the destination address field within the SN's local domain. The CN's EID is only allowed to use within the CN's local domain. After receive the reply, the LMC of the GW updates its flow table entries, and it sends the reply to the s_AR. Then the s_AR sends a command to the OFS for updating its flow table entries. Figure 2 shows the registration process and packet forwarding process outside of the local domain.

4 Conclusion and Future Works

In this paper, we propose an SDN-based ID-LOC separation architecture for IoT networks. We focus on how to provide scalability without any modification of the end nodes in IoT network, and try to overcome the shortcomings caused by the binding of IP addressing system. As a result, the paper introduces the basic architecture and describe how to generate the EIDs and LOCs. Also, we illustrate operating processes using the network architecture by dividing into the local domain and IP domain separately. The processes of registration, packet forwarding, and route optimization for the IoT devices are introduced. In the proposed architecture, hosts need not to participate in the signaling procedure, and do not consider any conversion procedure between EIDs and LOCs.

For the future works, the proposed scheme needs to be simulated or tested for accurate analysis and evaluation by comparing existing schemes in various conditions.

References

1. Moskowitz R, Nikander P, Jokela P, Henderson T (2008) Host identity protocol, IETF, RFC 5201
2. Nordmark E, Bagnulo M (2009) Shim6: level 3 multihoming shim protocol for IPv6, IETF, RFC 5533
3. Farinacci D, Fuller V, Meyer D, Lewis D (2013) The locator/ID separation protocol (LISP), IETF, RFC 6830
4. Luo H, Qin Y, Zhang H (2009) A DHT-based identifier-to-locator mapping approach for a scalable internet. *IEEE Trans Parallel Distrib Syst* 20(12)
5. Ratnasamy S, Francis P, Handley M, Karp R, Shenker S (2001) A scalable content-addressable network. In: *SIGCOMM'01*
6. ONF White Paper, Software-defined networking: the new norm for networks. <https://www.opennetworking.org/images/stories/downloads/sdnresources/white-papers/wp-sdn-newnorm.pdf>
7. Open Networking Foundation (2015) OpenFlow switch specification version 1.5.1, TS-025
8. Kushalnagar N, Montenegro G, Schumacher C (2007) IPv6 over low-power wireless personal area networks (6LoWPANs): overview, assumptions, problem statement, and goals, IETF, RFC 4919
9. IEEE Computer Society (2003) IEEE Std. 802.15.4-2003
10. Deering S, Hinden R (1998) Internet protocol, Version 6(IPv6), IETF, RFC 2460

An Efficient Disposition for Wrist-Worn Device Usage Time Expansion in Wearable Computing Environment



Jong Won Lee, Ji Su Park, Hyeong Geun Kim, and Jin Gon Shon

Abstract Wrist-worn devices have merits of usability such as glanceability, faster reaction time, and low access time. However, usability is restricted by difficulty from mid-air interaction which can cause fatigue and muscle damage. Traditional disposition for wrist-worn users considers only interaction between devices and users who wear devices outside of the wrist. In this paper, we propose efficient disposition (ED) which can expand the usage time for long time use. ED has two disposition for glanceability and long-time use. The former turns the power on all the time for time, blood pressure, heart rate, and blood sugar confirming with low energy consumption. The latter uses for text entry, viewing pictures and videos and so on. Consequently, ED can hold and sustain wrist-worn devices for a long time, therefore it enables users to enjoy better usability.

Keywords Disposition · Interaction · Wrist-worn device · Smartwatch · Wearable computer

1 Introduction

Wrist-worn devices such as smartwatch and fitness tracker have become commonplace in activity daily living [1]. Wrist-worn devices have advantages over other

J. W. Lee · H. G. Kim · J. G. Shon (✉)

Department of Computer Science, Graduate School, Korea National Open University, Seoul, South Korea

e-mail: jgshon@knou.ac.kr

J. W. Lee

e-mail: neuronet@knou.ac.kr

H. G. Kim

e-mail: hgrikim@knou.ac.kr

J. S. Park

Convergence Institute, Dongguk University, Seoul, South Korea

e-mail: bluejisu@dgu.edu

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_49

devices which are faster reaction time and access time [2]. However, their usability is restricted by a small touchscreen, brief access time, and glanceability. Then, they do not use as stand-alone devices, certainly use with other devices such as smartphone and desktop [1, 3].

As Wrist-worn device users increasingly adopt, interactions between user and device become more important, then many researchers create new interaction to support their needs. These interactions must be designed to consider not only at the moment of use but also for periods of use. Users take several poses in a situation that users use wrist-worn devices. Disposition indicates both given user poses and physical relations between user and device. Researchers must find out efficient disposition for users [4].

Users mainly hold weights such as the forearm, upper extremity and/or wrist-worn device in the air when they interact with their wrist-worn devices. That is mid-air interaction. It can evoke fatigue and feeling of heaviness in their forearm or upper extremity the so-called gorilla arm syndrome. In this paper, we quantify perceived exertion with self-report Borg CD10 scale based on muscle sensation which is 12 point scale from 0 to 10. It is a category ratio scale to assess perceived exertions [5].

In Only 3 min demonstrates disposition between smartwatch on the wrist and user's poses such as sitting and standing. In their pose, they perform three different input tasks such as touch, dwell and swipe. Then, researchers measured the perceived exertion change of users who feel somewhat strong in their arm. They find out the upper bound of time to use smartwatch, most perceived exertion task, and limits of smartwatch application and user interfaces through perceived exertion study [2]. However, In Only 3 min does not consider other pose and physical relations between user and smartwatch. In particular, users can wear a smartwatch on the inside of their wrist or on the radial aspect of it.

Despite developments of wrist-worn devices, there is insufficient to consider various dispositions between users and devices. Traditional dispositions for wrist-worn devices are faced outside of wrist then users flex the shoulder and elbow, then pronate their forearm for interaction their devices. So, they cannot reduce fatigue and increase the time of use. Therefore, we propose an efficient disposition for wrist-worn device users.

2 Design of Fatigue Reduced Disposition

ED is one of the dispositions of wrist-worn devices in upper extremities that it can reduce fatigue and then increase the time of use. ED is consist of disposition for glanceability and disposition for long time use. Figure 1 shows the design of ED.

Disposition for glanceability is placed wrist-worn devices on the lateral side of forearms. This site is placed in the middle range of motion between pronation and supination in their forearms during interactions between users and devices with shoulder flexion and elbow flexion. In this range, pronator muscles and supinator muscles do not contract maximally then they feel less fatigued. Moreover, disposition

Fig. 1 Design of FRD



for glanceability is interactions between users and the lateral side of forearms which has narrow space. Hence, they can be used to check the time, blood pressure, heart rate, blood sugar and so on without using their other hand.

Disposition for long time use is placed wrist-worn devices on the inside of forearms. This site is placed in the end range of motion between pronation and supination in their forearms during interactions between users and devices with shoulder flexion and elbow flexion. In this range, elbow flexor muscles contract mainly which is the most powerful muscle of elbow joint then it can sustain heavyweights for a long time comparing other muscles around the elbow joint. So, they feel less fatigued for a long time of use. Moreover, this disposition protects wrist-worn devices against damages from external force and easy to use in a situation of holding on to something. Figure 2 shows the disposition for glanceability and disposition for long time use.

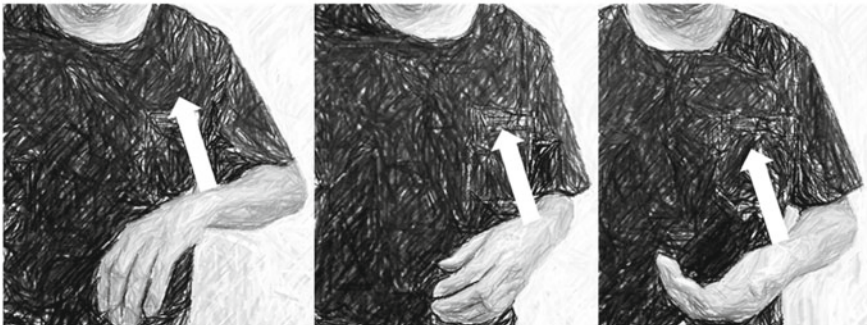


Fig. 2 Various disposition between users and devices such as in only 3 min (*left*), disposition for glanceability (*middle*), and disposition for long time use (*right*)

3 Performance Evaluation

3.1 Experimental Design

Thirty participants for the experiment have experience of smartwatch which varied from non-users to power user. They were 18 male (60%) and 12 female (40%) with average age of 29.83 ± 9.68 . We compare perceived exertions among various disposition for wrist-worn device users. Among various input tasks, we choose the touch input task because text entry is a key function of smartwatch app for interaction with other users [6]. So, participants perform touch input task with a Xiaomi Mi band 2 with different disposition for 300–330 s as shown in Fig. 2. These dispositions are similar to wear a watch on the wrist in a standing position with arm raised. We evaluate self-reported perceived exertions of participants with Borg CR 10 scale. Participants provide their perceived exertion by word of mouth then researchers notate their scale once every 30 s until their perceived exertion reaches “somewhat strong” which means 4 on the Borg CR10 scale. After that, they have a rest for more than 5 min. Statistical analysis is performed using SPSS version 18 (PASW) for windows. Paired t-test with the double-blind method was used and the order of disposition to evaluate was randomly arranged.

3.2 Experimental Result

Performances in a sitting position have several problems which cannot evaluate the influence of armrest height, the quantity of support, and the position of upper extremity and trunk. Moreover, it takes too long to evaluate perceived exertion in the experiment. For the reason, we perform experiment to only four participants for the understanding tendency in the sitting position as shown in Fig. 3.

Figure 4 shows the boxplots of the perceived exertion scale over time. Borg scale of participants correlates with time, then there is a significant influence of time on perceived exertion statistically. Thus, perceived exertion of wrist-worn devices user increased proportionally to time.

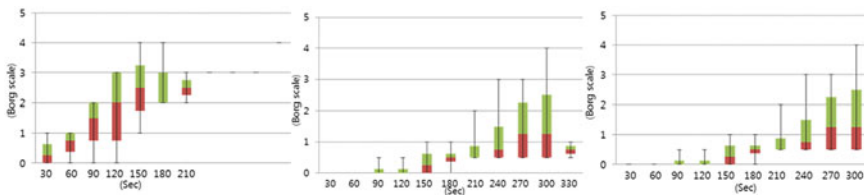


Fig. 3 Boxplot for borg scale for participants in sitting position for all three interactions: in only 3 min (left), disposition for glanceability (middle), and disposition for long time use (right)

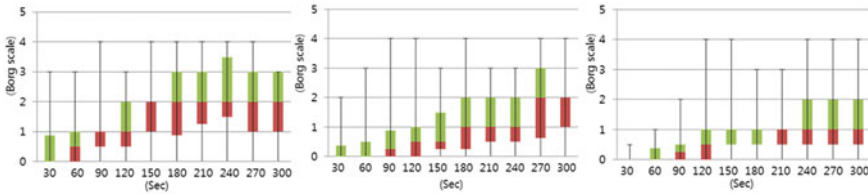


Fig. 4 Boxplot for borg scale for participants in standing position for all three interactions: in only 3 min (*left*), disposition for glanceability (*middle*), and disposition for long time use (*right*)

The result of the experiment on perceived exertion of various disposition for wrist-worn device users was as follow; the average time to 4 on the Borg CR10 scale of In Only 3 min was 230.5 ± 78.26 s, the average time to 4 on the Borg CR10 scale of disposition for glanceability (Glance-Disp) was 271.67 ± 64.01 s, and the average time to 4 on the Borg CR10 scale of disposition for long time use (Long Time Use-Dispo) was 280.17 ± 49.59 s as shown in Fig. 5. Disposition for glanceability was on average 41.17 ± 69.62 s (<0.05) longer than In Only 3 min. Also, disposition for long time use was on average 49.67 ± 75.04 s (<0.05) longer than In Only 3 min. Whereas disposition for long time use was on average 8.50 ± 42.73 s (>0.05) longer than disposition for glanceability that is not statistically significant. Figure 5 illustrates the performance evaluation and the results are given in Table 1.

Fig. 5 Result of performance evaluation about perceived exertion among various disposition

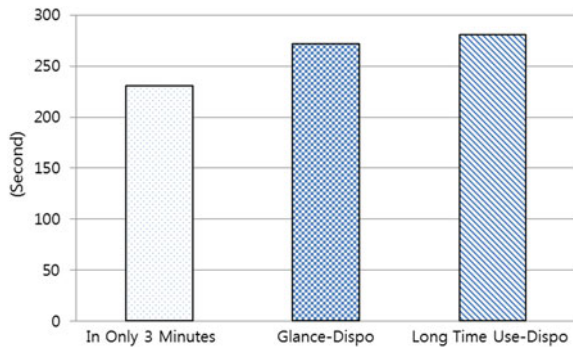


Table 1 Results of experiments

	M ± SD	p-value
Glance-dispo— in only 3 min	41.17 ± 69.62	0.003
Long time use-dispo— in only 3 min	49.67 ± 75.04	0.001
Long time use-dispo—glance-dispo	8.50 ± 42.73	0.285

4 Conclusion

An efficient disposition is one of disposition of upper extremity for interaction between users and device which is consist of disposition for glanceability and disposition for long time use at the wrist. ED has the following contributions. It can confirm quickly short information and message without using their other hand through the radial aspect of their wrist. It also reduces muscle fatigue on forearms and upper extremity and risk of musculoskeletal disorder due to biomechanical overload through inside and radial aspect of their wrist. Thus, it can prolong the usage time when they interact with wearable computer especially at wrist and forearm. According to the performance comparison, ED can reduce rated perceived exertion then it can expand the usage time. Consequently, it can enjoy better usability than before in extended interactions by considering efficient dispositions.

References

1. Rushil K, Nikola B, Kent L (2018) In only 3 minutes: perceived exertion limits of smartwatch use. In: Proceedings of the 2018 ACM international symposium on wearable computers. ACM, New York, pp 208–211
2. Daniel LA, James RC, Kent L, Thad ES, Nirmal P (2008) Quickdraw: the impact of mobility and on-body placement on device access time. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 219–222
3. Kyung-Taek L, Hyoseok Y, Youn-Sung L (2018) Implementation of smartwatch user interface using machine learning based motion recognition. In: 2018 international conference on information networking (ICOIN). IEEE, New York, pp 807–809
4. Kent L, Halley P (2014) The multiple dispositions of on-body and wearable devices. *IEEE Pervasive Comput* 13(4):24–31. IEEE
5. Sujin J, Wolfgang S, Satyajit A, Karthik R (2017) Modeling cumulative arm fatigue in mid-air interaction based on perceived exertion and kinetics of arm motion. In: Proceedings of the 2017 CHI conference on human factors in computing systems. ACM, New York, pp 3328–3339
6. Andreas K, Mark D (2014) Text input on a smart watch. *IEEE Pervasive Comput* 13(4):50–58. IEEE

Preserving Sustainability for Mission-Oriented Cyber-Physical Systems Collaboration



Horn Daneth, Nazakat Ali, and Jang-Eui Hong

Abstract Cyber-Physical Systems (CPSs) were become a widely used system in recent smart era, and are also becoming more complex and diverse due to increasing their application areas. The interaction and collaboration among CPSs to provide services or achieve missions to customers are very important because their malfunctioning may cause serious accidents. This paper proposes a framework for preserving sustainability to accomplish the given mission of collaborative CPSs. Especially, our framework firstly monitor the violation of safety that can occur in the cooperation among the systems and then validate the coordination of each task execution for mission-oriented collaboration. Our framework can support building interconnected systems for highly assured services.

Keywords Cyber physical systems · Collaborative CPSs · Mission-oriented systems · Sustainability framework · System safety

1 Introduction

Cyber-Physical Systems (CPSs) are co-engineered interacting networks of physical and computational components [1]. These systems will provide the foundation for our social infrastructure, form the basis of emerging and future smart services, and improve our quality of life in many areas. Increasingly, such systems will be everywhere, from intelligent robots to autonomous vehicles or to smart factory.

Along with the advance in ICT technology, more complex and compound services will be required more and more through the collaboration of CPSs [2, 3]. For

H. Daneth · N. Ali · J.-E. Hong (✉)

Department of Computer Science, Chungbuk National University, Cheongju 28644, South Korea
e-mail: jehong@chungbuk.ac.kr

H. Daneth

e-mail: horndaneth@selab.cbnu.ac.kr

N. Ali

e-mail: nazakatali@selab.cbnu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_50

363

example, in order to recover disasters such as earthquakes and tsunamis, a variety of autonomous robots—firefighting robot, lifesaving robot, dismantling robot, etc.—have to collaborate to complete their common mission. Whereas, each robot has been developed independently in order to perform their specific task. Therefore, a systematic coordination is required when they interact with each other to achieve a mission because the behaviors of the dismantling robot can interfere with the behaviors of lifesaving robot [4].

It is very critical issue to ensure that safe collaboration among CPSs must be preserved continuously until the mission is achieved because failure of the coordination may cause damage, loss, injury and/or deaths [5]. We define a framework for preserving safe and sustainable collaboration among CPSs. Our framework, SuCoF (Sustainable Collaboration Framework for CPSs) monitors the behaviors of each CPS and coordinates the whole members for safe operations when independent CPSs were connected each other in order to complete a common mission. Our framework, SuCoF proactively supports the dynamic reconfiguration of CPS platoon to preserve sustainable collaboration because it needs to control and manage the CPSs through continuous monitoring at runtime to ensure safety against unexpected situations that did not be considered in the development of CPS.

Similar researches have also been conducted on coordination and control of mutual collaboration. Azfar [6]’s work proposed a realistic methodology for collaboration between human and CPSs in industry, and Adam [7]’s work proposed human–machine interface considering model-based development in medical domain. However, the studies did not concern about CPSs collaboration without human intervention. In our research, we propose a solution for the CPSs collaboration problems (i.e., “dynamic reconfiguration” and “collaboration without human intervention”) in terms of SoS (System of Systems).

This paper is organized as follows: Sect. 2 defines our SuCoF framework and its components in details. The techniques and algorithms for safe and sustainable collaboration are explained in Sect. 3. In Sect. 4, a case study was investigated to validate our framework. Section 5 concludes our work and suggests further research work.

2 Sustainable Collaboration Frameworks

Our proposing framework for supporting safe and sustainable collaboration among CPSs is shown Fig. 1, which consists of major three parts; CPS_SM, PLTN_SL, and PLTN_LD which are CPS service manager, Platoon subleader, and Platoon leader, respectively. The CPS_SM has capable of check whether their states may be able to perform the given mission and can decide about reaching final state, and the PLTN_SL is responsible for performing the mission given from PLTN_LD in its management platooning space. The PLTN_LD has the same responsibility (i.e., same components) with the PLTN_SL. But PLTN_LD only issues a mission to PLTN_SL and receive mission list from external agent.

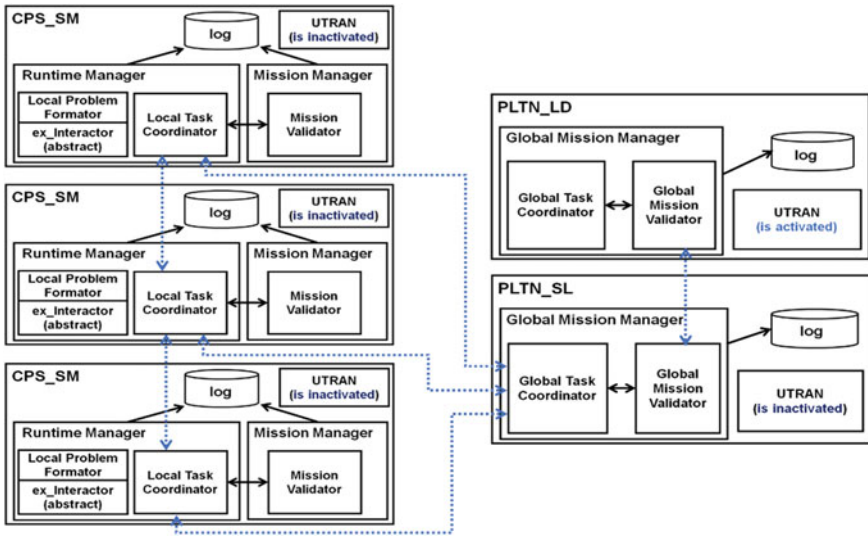


Fig. 1 Overall architecture of the framework (SuCoF) supporting safe and sustainable collaboration among cyber-physical systems

Each part has some components to perform its responsibility. The RTM (Runtime Manager) checks for any abnormal state that may occur while the CPS_SM performs each task. The MM (Mission Manager) checks that the CPS_SM has received a valid mission from the PLTN_SL and that whether the result of task coordination violates their mission execution. The GTC (Global Task Coordinator) executes global mission, i.e., coordinates tasks of the CPS_SMs to fulfill the given mission. Through the GMV (Global Mission Validator) of both the PLTN_LD and the PLTN_SL, the mission-list received from agent can be mapped to CPS_SM mission (i.e., set of tasks of CPS_SM).

Through this framework, the PLTN_LD receives missions from external agent, send them to PLTN_SL that delivers the missions to platoon members, and oversee global orchestration. The CPS_SMs as platoon members perform tasks that compose a mission by checking whether the given mission can be performed in the current state. These behaviors allow the given mission to be performed safely and sufficiently in our framework.

3 Preserving Sustainability

3.1 Potential Hazards

The potential hazards that can occur during CPS collaboration can vary. As a representative example, a platooning system can be considered, as shown in (a) of Fig. 2, which is configured with a leader L, three subleaders S1, S2, and S3, and platoon members within subleader boundaries P1, P2, and P3. In collaborating situation, if the subleader S1 malfunctions and cannot communicate with the members of the P1 range as shown in (b) of Fig. 2, the members of the platoon group P1 lose the coordinator for collaboration. In order to continue to collaborate in this situation, new subleader must be quickly selected and continued to perform their mission within the P1 group.

In addition to these situations, there are potential hazards of collaboration that can be considered when a leader fails or when a member of platoon group leaves or moves to another group.

3.2 Sustainable Collaboration

When the components in the framework SuCoF basically communicates each other based on IEEE 802.11-based VANETs (Vehicular Ad-hoc Networks [8]) (where, data packet delivery ratio, control packet overhead, throughput, packet drop fraction was not considered), in order to resolve the abnormal state, as shown in Fig. 2, the following subleader (S) election algorithm can be used to accomplish the mission through continuous collaboration within the platoon group P_k .

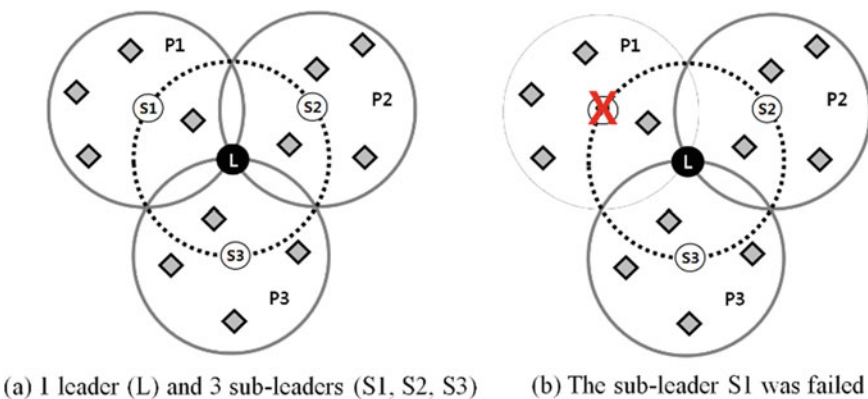


Fig. 2 Collaboration through subleaders in the interaction space: **a** configuration of normal collaboration, **b** abnormal state due to the failure of the subleader S1

Algorithm ElectSubleader()
 $p_i \in P$ sends Check(S_i) to $\forall p_j \in P_k$;
 $\forall p_j$ send Call(S_i) to subleader S_i ;
 Wait for OK(S_i) from S_i ;
if ($\forall p_j$ did not receive OK(S_i)) **then**
 $\forall p_j$ tests ping(L) to leader L;
 For ($\forall p_j$ that receive Ping.Ack)
 Get ping time for L;
 Send pingTime(p_i, L) to $\forall p_j \in P_k$;
 Find a p_j that has Minimum {pingTime};
 p_i sends newSubleader(p_j, P_k) to L;
 L sends corresponding mission list to p_j ;
 $\forall p_j \in P_k$ set the p_j as new S_i ;
 The p_j generates member list for P;
End Algorithm

Above algorithm starts when any platoon member p_i did not receive commitment for its finished task from subleader, and selects new subleader that has minimum turn-around time value of the ‘ping test’ for leader. This allows selecting a collaborative coordinator who can continue to perform a mission from the dead state that could be caused by an unexpected subleader failure.

4 Application: People Rescuing Mission

4.1 Hazardous Scenario

The behaviors of autonomous robots participating in people rescuing from disaster can be represented with state transition diagrams, as shown in Fig. 3.

The Fig. 3a represents the behavior of the robot that dismantles and removes the objects or obstacles, the (b) represents the behavior of detecting life by irradiating infrared beam, and the (c) is for the behavior of performing first-aid for patient.

Even if the following situations occur when the above autonomous robots collaborate in unforeseen circumstances, the given mission must be maintained at a level that can be tolerated.

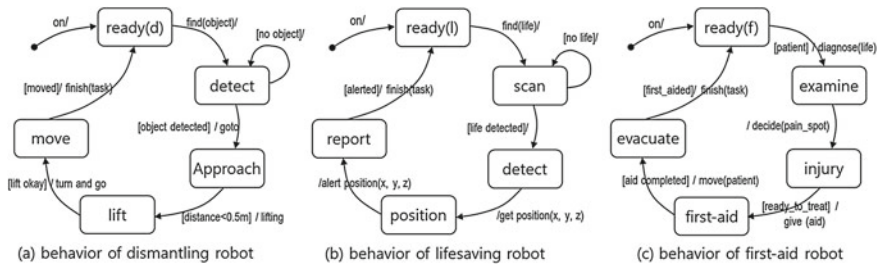


Fig. 3 State transition diagrams of autonomous robots for disaster recovery

Subleader failure. One of participating robots in mission space should be elected to subleader when the subleader did not work. The new subleader performs its roles continuously after receiving the mission list from leader.

Member failure. Because the subleader cannot receive a message of task completion from a member robot that is failed, the subleader deletes the tasks of the failed robot from mission list, and re-serializes the remaining tasks in the list. The subleader also reports this information to the leader.

4.2 Global Task Coordination

When the robots in Fig. 3 collaborate, there is a subleader that manages them. The subleader maintains the task list for their collaboration by serializing or parallelizing the tasks. The mission to be achieved by the subleader can be written in the Wright specification language [9] as shown in Fig. 4.

The Wright specification in Fig. 4 includes the phrase ‘safety contract’ to indicate that the interface should be satisfied when independent robots interact with each other [10]. The safety contract defines Sa (strong contract), Sg (strong assumption)

```

SuCoF RescuingPeopleSystem
CPS_SM DismantlingRobot =
  port ownData = transmit → ownData □ ξ
  port otherRemoteData = receive → otherRemoteData □ ξ
  computation = TransmitRemoteData
  where
    TransmitRemoteData = internalComputation → ownData.transmit → (ReceiveRemoteData □ ξ)
    ReceiveRemoteData = otherRemoteData.receive → TransmitRemoteData □ ReceiveRemoteData □ ξ
  computation = ApproachingtoObject
  where
    ApproachingtoObject = goforward → ownData.transmit → (ReceiveRemoteData □ ξ)
    ReceiveRemoteData = otherRemoteData.receive → (approachtoObject□ξ)
  computation = LiftingObject
  where
    LiftingObject = GripObject → LiftingObject → RemoveObject → ownData.transmit → (ReceiveRemoteData □ ξ)
    ReceiveRemoteData = otherRemoteData.receive → (LiftingObject□ξ)
CPS_SM LifeSavingRobot =
  port ownData = transmit → ownData □ ξ
  :
CPS_SM FirstAidRobot =
  port ownData = transmit → ownData □ ξ
  :
Connector CPS-connector =
  role dismantler = (ready; detect; approach; lift; move) □ ξ
  role saver = (ready; scan; detect; position; report) □ ξ
  role firstaider = (ready; examine; injury; first-aid; evacuate) □ ξ
  glue = saver.report → dismantler.approach → glue □ dismantler.move → firstaider.examine → glue □ ξ
Configurator PLTN_SL =
  configuration [dynamic configuration among participating robots and attachments]
  coordination [task coordination in serial | parallel processing]
Instances
  D: DismatlingRobot; L: LifeSavingRobot; F: FirstAidRobot; SL: PLTN_SL; L: PLTN_LD
Safety Contracts
  SC = < Sa1, Sg1, { < Wa1, Wg1 >, ..., < Wan, Wgn > } > □ < Sa2, Sg2, { < Wa1, Wg1 >, ..., < Wan, Wgn > } > □ < Sa3, ..
End RescuingPeopleSystem

```

Fig. 4 Wright specification for collaborating autonomous robots in disaster situation

and Wa (weak contract) and Wg (weak assumption) that CPSs must meet in the interactions that occur during collaboration, which coordinates them to complete their tasks through the normal execution of their assigned tasks.

The phrase ‘Configurator’ in Fig. 4 defines the role of the subleader; the element ‘configuration’ includes the number and roles of participating robots, and the element ‘coordination’ contains a list of all tasks that the participating robots to be performed, and used to control for transitions of behavioral or protocol states.

In order to verify the Wright specification for safe and sustainable collaboration, the specification is converted to CPS specification using the wr2fdr tool, and the FDR2 tool [11] can verify the functional correctness of the collaboration. The FDR2 tool provides the information of the number of states and transitions, time (in second) to verify deadlock-freeness, time (logical time) to finish the given mission. In the simulation of our case, the number of states and transitions are 70 and 237, respectively. The time to find deadlock (or deadlock-freeness) is required to 0.23 s.

5 Conclusion and Further Work

This paper proposes a framework, SuCoF that enables a CPS with independent functions to perform its tasks safely and sustainably when given a mission that must be achieved through mutual collaboration. This framework consists of leader, subleaders, and members managed by the subleader in order to carry out a given collaboration mission, also provide those functions of runtime manager and mission manager in order to ensure the safety and sustainability of their collaboration.

We identified potential hazards that might occur in the framework and suggested algorithms and method to solve them. In addition, to verify the effectiveness of the proposed method, we introduced the Wright specification with an extension to specify the safety contract, and the functional simulation was made based on the FDR2 tool.

Our future work will increase the practicality of the proposed framework by deriving, modeling and verifying hazards that can occur during collaboration at a more detailed instruction level.

Acknowledgements This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT (NRF-2017M3C4A7066479).

References

1. Alur R (2015) Principles of cyber-physical systems. MIT Press
2. Herterich MM, Uebernickel F et al (2015) The impact of cyber-physical systems on industrial services in manufacturing. In: 7th industrial product-service systems conference, vol 30, pp 323–328
3. Engell S (2014) Cyber-physical systems of systems—definition and core research and innovation areas. European roadmap on research and innovation in engineering and management of cyber-physical systems of systems, p 111
4. Platbrood F, Gornemann O (2017) Safe robotics—safety in collaborative robot systems. SICK Sensor Intelligence, pp 1–8
5. Pallottino L, Scordio VG et al (2007) Decentralized cooperative policy for conflict resolution in multi-vehicle systems. *IEEE Trans Rob* 23:1170–1183
6. Azfar K, Kirisci PT et al (2016) A methodology to develop collaborative robotic cyber physical systems for production environments. *Logist Res* 9(23):1–15
7. Adam N (2010) Workshop on future directions in cyber-physical systems security. Report on workshop organized by Department of Homeland Security
8. Benslimane A, Taleb T et al (2011) Dynamic clustering-based adaptive mobile gateway management in integrated VANET-3G heterogeneous wireless networks. *IEEE Commun* 29(3)
9. Allen R, Douence R, Garlen D (1998) Specifying dynamism in software architectures. *Fundamental approaches to software engineering*, pp 21–37
10. Medawar S, Scholle D, Sljivo I (2017) Cooperative safety critical CPS platooning in SafeCOP. In: The 6th Mediterranean conference on embedded computing, June
11. Miyazawa A, Ribeiro P et al (2019) RoboChart: modelling and verification of the functional behaviour of robotic applications. *Softw Syst Model* 18:3097–3149

A 3D Object Segmentation Method Using CCL Algorithm for LiDAR Point Cloud



Yifei Tian, Wei Song, Jinming Liu, and Simon James Fong

Abstract Environment perception and analysis are essential parts of automatic ground vehicles (UGVs) to implement smart driving-decision making. To sense dynamic environment information, light detection and ranging (LiDAR) are equipped on UGVs to collect high-precision 3-dimension point cloud. Because of the unstructured and inhomogeneous characteristics of LiDAR point cloud, the fast and accurate analysis of point cloud is hard to achieve under UGVs' driving. Most UGVs' autonomous applications, such as object extraction, terrain perception, and traversable path recognition, face technology bottleneck in both process speed and analysis precision. In these applications, object segmentation result is a fundamental information support, which influence the subsequent processes to a large extent in both accuracy and efficiency performance. This paper proposed a novel object segmenting algorithm named 3D connected component label (3D-CCL) to divide full point cloud into subdivision point clouds of individual local object space. The object segmenting result provide a series of basic point clouds of different obstacle models, which benefits for the environment perception and decision making for UGV.

1 Introduction

Fast and precision object segmentation of raw environment datasets is a necessarily requirement in a large number of UGVs' autonomous perception applications, including obstacle prediction, traversable road detection, and dynamic-static scene segmentation [1–3]. Accurate segmenting results can maintain a reliable input for the consistent urban environment model reconstruction, which always consists of trees, pedestrian, building, telephone pole, and other city infrastructures [4]. The object

W. Song (✉) · J. Liu
Department of Digital Media Technology, North China University of Technology, Beijing CO
80305, China
e-mail: sw@ncut.edu

Y. Tian · S. J. Fong
Department of Computer and Information Science, University of Macau, Macau CO 80523, China

segmentation and extraction from the dynamic and complex terrain information dataset became a popular research point in autonomous driving domain [5].

To percept environment information, digital cameras, depth cameras (e.g., Kinect), light detection and ranging (LiDAR), and other sensors are widely equipped on UGVs [6]. Compared with LiDAR, the overwhelmingly disadvantages of digital cameras and depth cameras are that sensitive to illumination changing and limited by range of view. 3-Dimension (3D) LiDAR is able to collect high-precision and large-scale point cloud of distance information from UGV's driving environment. A large amount of LiDAR points are sensed within a short time, which make the sensor suitable to be carried on UGVs' to face the changeable environment.

Point cloud collected by LiDAR are sorted unstructured and inhomogeneous so that the analyzing process of LiDAR point cloud is computational complexing and time consuming. Besides, point cloud always distributed sparsely and without symmetrical density, which lead to difficult pre-process for points' to seek their neighborhoods by using a predefined threshold. In several traditional neighbor points seeking methods, a radius r (distance-fixed) or constant k (density-adaptive) are defined to restrict their searching area. However, applying these methods in LiDAR points to search neighbor points are impractical according to the disordered arrangement, which burden massive computation to detect neighbors by recursive.

This paper proposed a 3D connected component labeling (3D-CCL) algorithm to segment full point cloud into corresponding local objects. In the proposed method, ground points are filtered out from the raw LiDAR point cloud as the first step before sequence process. Then, map all the non-ground points into a predefined 3D grid box, consisting of given-sized cube units. Through clustering valid neighborhood cubes, adjacent 3D points mapped in the clustered cubes are grouped together based on the 3D-CCL algorithm.

The remainder of this paper is organized as follows. Section 2 describes the 3D-CCL algorithm by using parallel computation. Section 3 illustrates the experiment results. Finally, Sect. 4 concludes this paper.

2 CCL Algorithm Interpretation

In the CCL algorithm, all the clustering processes are only employed on valid grid cells defined as above. After filtering invalid grid cells, valid grid cells obtain their corresponding index values I according to the relative locations i, j , and k in the flag box. Thus, each I is a unique value in the flag box to identify the valid grid cell, and at the same time the index values of each invalid cell is set as null.

To mark the grid cells belonging to a same obstacle, a same and unique label is assigned for these grid cells as their identification. For convenience, all the grid cells belong to a same object are assigned the minimum index values as their unanimous labels L_i , where the minimum index is searching from the grid cells' index values belonging to a same object. For implementing the label updating target, each cell's label is set as the minimum index value from its neighbor cells' index values as the

first step. Through updating index values of neighbor cells, all the cells' labels are set as their goal labels by several times of label updating iteration.

To detect whether the label of neighbor grid cells are the minimum one, a 3D descriptor are utilized to define the searching areas of grid cell unit in flag box. When the searching step equal one, the center cell contains six neighbor cells in six different directions as up, down, left, right, front, and back. According to the 3D structure of the example descriptor, only one neighbor cell in each direction is covered by the descriptor because of the searching step is set as one.

Every grid cells in the flag box are traversed by using the descriptor through several iterations until valid cells' labels are not change anymore. In each minimum label searching, the center cell is required to be a valid one and also only valid neighbor cells would take part in the searching process. Just like a kernel or an operator in image domain, our descriptor is utilized to process the cells' labels in the rasterized 3D flag box. In each local searching, when the labels L_i in the descriptor are not same, all the cells' labels are updated as the smallest one. As shown in Eq. (1), the searching area U of a general descriptor is defined as follow. Variable i and i' are the index values of the center grid cell and the neighbor grid cells, respectively. The search step S is a three dimension vector consisting of s_x, s_y, s_z , where all the s_x, s_y, s_z equal 1 in the example descriptor.

$$U(I) = \{(I') = (I + \Delta S) | (-S \leq \Delta S \leq S)\} \quad (1)$$

$$I' = \{(i', j', k') = (i, j, k) + \Delta S\} \quad (2)$$

$$\Delta S = \{(\Delta x, \Delta y, \Delta z) | (-s_x \leq \Delta x \leq s_x), (-s_y \leq \Delta y \leq s_y), (-s_z \leq \Delta z \leq s_z)\} \quad (3)$$

In the Eq. (2), the three dimension vector ΔS are contains $\Delta x, \Delta y, \Delta z$ values, whose variables' ranges are demonstrated in Eq. (3). To find the minimum index value in the defined descriptor, the label are updated according to the Eq. (4).

$$\begin{aligned} L(I) &= \min(\forall L(I')) \\ (I') &\in U(I) \end{aligned} \quad (4)$$

After executing the label updating process, the grid cells belong to a same group own a same label as their identification through several iterations. When the cells' labels of in the flag box are not changed anymore, the connected cell clustering are considered as finished thereafter the grid cells with same label are considered as a same object. Based on the clustering result in flag box, the labels are inversely mapped onto the 3D point cloud so that points carried a same label are considered as belonging to a same object.

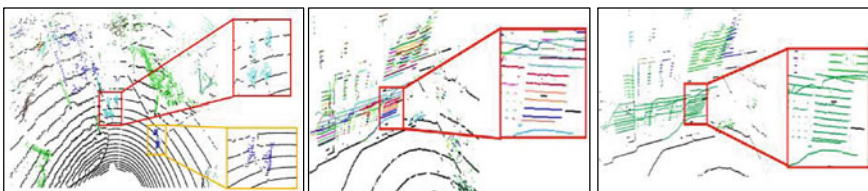
3 Experiment

Our system used a Velodyne LiDAR sensor to precept environment information around unmanned vehicle, in which a laptop computer is used to store the sensed point cloud. The distance of the collected point cloud cover 70 m within less than 2 cm' measure error. The perception angle of the 3D LiDAR contains 360° in horizontal and -30.67° to $+10.67^\circ$ in vertical directions. Each frame contains around 96,000 high-precision points, which cause large computational consumption in segmenting process. The segmenting program was executed on a 2.80 GHz Intel® Core™ i7-7700HQ CPU computer with an Nvidia GT 1500Ti graphics card and 8 GB RAM. The point cloud datasets are collected in an urban street in Beijing.

Figure 1 demonstrate three segmenting situations, including under segmented, over segmented, and normal segmented situations shown in (a), (b), and (c), respectively. When a resolution is set as a relatively big one, under segmented problem would exist in segmenting process. Figure 1a represents the under segmented situation, in which several pedestrians as clustered into one group. Figure 1b shown the over segmented situation, in which a wall is divided into several different pole-like objects. The normal segmented result is shown in Fig. 1c, where the wall object is segmented correctly and rendered in shallow green.

The segmenting results using the proposed CCL algorithm based on different resolution are illustrated in Fig. 2, where individual objects are rendered in different colors. The experiment used 4 different resolutions to segment 4 environment frames consisting of large size of LiDAR point cloud. Figures in same column represent segmenting results in a same environment based on different resolutions. The resolutions in the 4 rows from up to down are 1.5, 2.0, 2.75, and 3.0, respectively. These figures show the conspicuous segmenting differences under different resolutions when comparing with the figures belonging to a same column.

In the first row, wall objects are not considered as a same object as shown in the left part of (a), (b). Tree objects are also segmented as different objects because their sparse branches and leaves in (b), (c), and (d). In second row, the over segmented problem in (e) (f) (g) (h) is slightly better than (a) (b) (c) (d) in the first row under. For example, the tree objects at bottom-right part of the (g) absolutely belong to two trees, but the point cloud around it is separated as different objects in (c). Comparing



(a) Under segmented result (b) Over segmented result (c) Normal segmented result

Fig. 1 Three segmented results under different resolution types

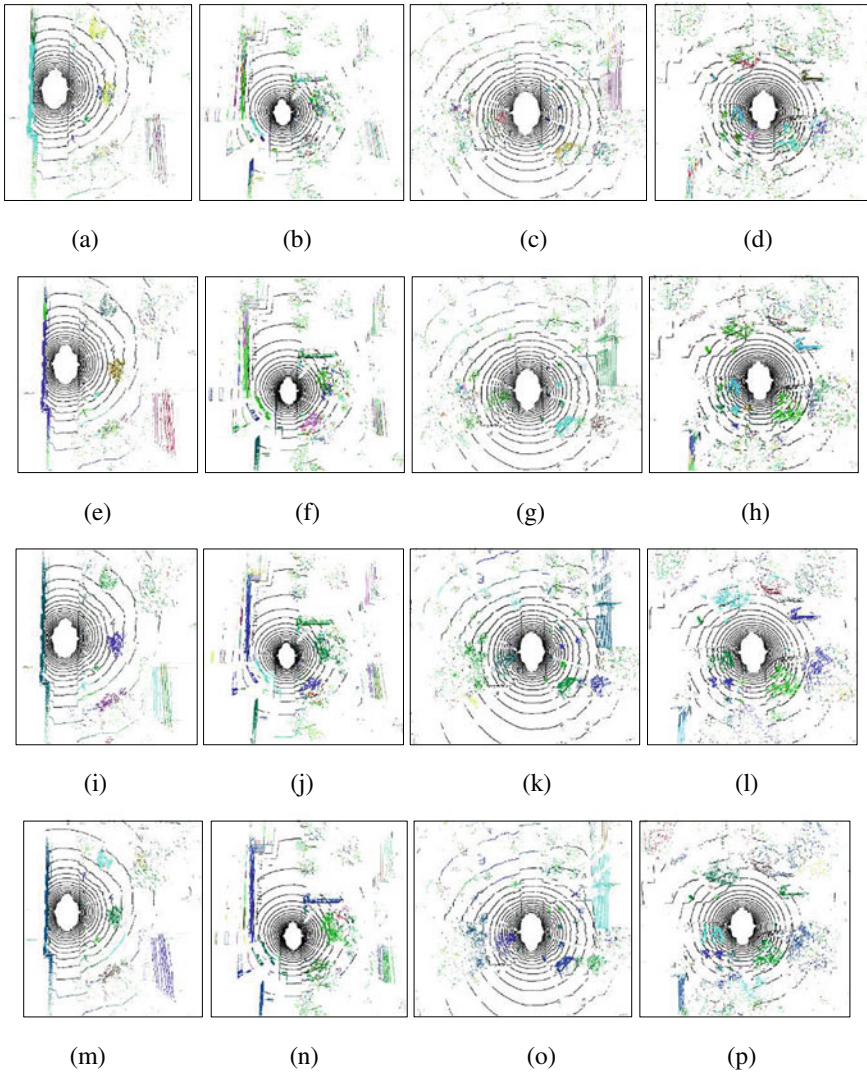


Fig. 2 Segmenting results in different frames and resolutions

the figures in third and forth row, the segmenting results in third row is better than that in forth row. For instance, point cloud in the bottom-right part of figure (n) are not belonging to same objects, which is segmented incorrectly as a same object rendered in dark green. Under segmented problem exist in the forth-row figures because the resolution in forth row is large than the suitable range.

To realize precision object segmenting, a suitable resolution is required in the proposed 3D-CCL algorithm. Because our datasets were collected in summer, the

leaves on both trees and bushes are mainly represented as lots of loose and unstructured sphere-like objects. Tree objects are always existed in our collected datasets, which is a big segmenting difficulty in our proposed segmentation system. Considering the segmenting results in Fig. 2, the resolution 2.0 and 2.75 is better than that of 1.5 and 3.0 by comparing the selected 4 frames. Only the segmenting results based on the 4 searching steps are not justifiable enough to decide which parameter is the most suitable one faced with our collected datasets.

4 Conclusion

In this paper, we developed a fast object segmentation algorithm to divide LiDAR point cloud into individual object point clouds in different local sub-spaces. Our proposed 3D-CCL algorithm can realize object segmentation in high speed performance and segmenting precision. Through comparing different segmenting resolution and searching step, the most suitable parameters obtained from our environment datasets is 2.75 and 1, respectively. Thus, segmentation result of our proposed algorithm is able to support efficient environment information for sequence applications required on UGVs, such as object recognition, obstacle avoidance, local path planning, and so on. In the future, we will do more research on these sequence applications for improving the automatic driving safety and smart path planning by using the segmenting result.

Acknowledgements This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program) (2019-0-01585) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), by National Natural Science Foundation of China (61503005), the Great Wall Scholar Program (CIT&TCD20190304, CIT&TCD20190305), Beijing Young Topnotch Talents Cultivation Program (CIT&TCD201904009), Research Project of Beijing Municipal Education Commission (No. KM201810009005), North China University of Technology “The Belt and Road” Countries Talent Training Base Project, and “Yuyou” Project of North China University of Technology.

References

1. Rozsa Z, Sziranyi T (2018) Obstacle prediction for automated guided vehicles based on point clouds measured. *IEEE Trans Intell Transp Syst* 19:2708–2720
2. Gu S, Lu T, Zhang Y et al (2018) 3-D LiDAR + monocular camera: an inverse-depth-induced fusion framework for urban road detection. *IEEE Trans Intell Veh* 3(3):351–360
3. Xiao Z, Dai B, Wu T et al (2017) Dense scene flow based coarse-to-fine rigid moving object detection for autonomous vehicle. *IEEE Access* 5:23492–23501
4. Hackel T, Wegner JD, Schindler K (2016) Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS Ann Photogramm Remote Sens Spat Inf Sci* III(3):177–184
5. Chen Y, Wang J, Xia R et al (2019) The visual object tracking algorithm research based on adaptive combination kernel. *J Ambient Intell Hum Comput*. <https://doi.org/10.1007/s12652-018-01171-4>.

6. Su T, Zhang S (2018) Multi-scale segmentation method based on binary merge tree and class label information. *IEEE Access* 6:17801–17816
7. Arnaud A, Gouiffès M, Ammi M (2018) On the fly plane detection and time consistency for indoor building wall recognition using a tablet equipped with a depth sensor. *IEEE Access* 6:17643–17652

A Real-Time Human Posture Recognition System Using Internet of Things (IoT) Based on LoRa Wireless Network



Wei Song, Jinqiao Liao, and Jinkun Han

Abstract Posture recognition technologies based on the Internet of Things (IoT) are widely required to care industry, such as fall detection of elder persons. In order to realize the low-power transformation of motion information in wide-area, Long Range (LoRa) is used in this paper to develop a human posture recognition system. The system is integrated by an mpu-9250 sensor, a LoRa Shield board and an Arduino Mega master control board, which collect human posture data and transmit them to the cloud server remotely. Combined with a random forest algorithm, real-time human posture movement data is carried out to recognize and classify human posture movement. The posture recognizing accuracy calculated by random forest algorithm is the higher than that of other classic machine learning algorithms. This way, our proposed real-time human posture recognition system is able to assist care industry to automatically monitor real-time posture situations of elder persons.

Keywords LoRa · IoT · Posture recognition · Random forest

1 Introduction

Monitoring elders through multiple sensors helps researchers conveniently understand both the elders' daily routines and occurrence probability of some unexpected situations. In addition, the monitoring elderly's life all time can effectively increase rescue efficiency if medical personnel notice the accident happening and support assistance timely. Besides, the human posture monitoring is a large part of daily

W. Song (✉) · J. Liao

North China University of Technology, No. 5 Jinyuanzhuang Road, Shijingshan District, Beijing 100-144, China

e-mail: sw@ncut.edu.cn

J. Liao

e-mail: liaojq1994@sina.com

J. Han

Department of Computer Science, Georgia State University, Atlanta, USA

e-mail: hjinkun1@student.gsu.edu

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_52

routine monitoring of elders' who living alone in their home. Therefore, the study of human posture recognition is of great significance for daily monitoring not only in care industry but also resident life domain.

At home and abroad, the recognition and research of human body posture are also widely concerned by researchers. Bao et al. [1] proposed a body posture study based on Kinect sensor, using which to identify behavioral data of the elderly. However, the detecting range of Kinect is limited, only parts of elderly's motion can be effectively detected. To avoid the limitation of detecting range, Zhang's team [2] proposed a rolling positioning system based on Micro-Electro-Mechanical System (MEMS) device to solve the previous problem. However, smartwatches require frequent charging, and the data collected by the triaxial accelerometer in MEMS only contains acceleration information in 3 axis. The nine-axis sensor named mpu-9250 that used in this paper still collect angular velocity and magnetometer information, which support more diverse data than smartwatches to assist human posture recognition [3]. Besides, LoRa is a wide area protocol with low power feature so that it help Internet of Things (IoT) system maintain abiding and extensive working situation [4]. In machine learning field, random forest algorithm is a common and efficient classifier to recognize human posture accurately and real-timely [5].

In this paper, a machine learning algorithm and low-power IoT devices are combined to construct real-time human posture recognition system. In the system, six kinds of human posture, including running, walking, standing, going downstairs, jumping, going upstairs, are recognized timely. The rest of this article is organized as follows. The second part introduces the application and realization of the proposed human posture recognition system. The third part gives the algorithm details and experimental results of this system. The fourth part is the conclusion and our future work.

2 Human Posture Recognition System

In this paper, a nine-axis attitude sensor (Mpu-9250) is used to collect the attitude data of human body. The sensor own several significant advantages, such as low power consumption, low price and high functional requirements. As shown in Fig. 1, the

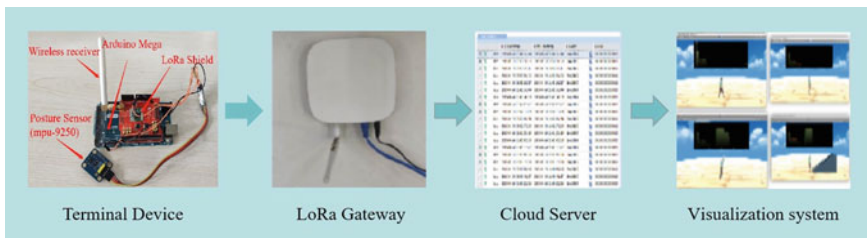


Fig. 1 Architecture diagram of the our proposed system

terminal device wirelessly sends the collected posture data to the gateway by Long Range (LoRa) Shield. The LoRa shield provides ultra-long range spread spectrum communication and high anti-interference functions with low electric consumption. The LoRa gateway preprocess and encrypt the received data and send the processed data to the cloud server for subsequent steps. In the cloud server, a random forest classifier is established and trained by using the collected data to recognize human postures. Based the classifying results, a visualization system visualizes these human posture data by six kinds of different animations for help users understanding current posture situations intuitively.

Random forest adopts random resampling technique and node random splitting technique to generate multiple decision trees and receive the final result through voting. First, N samples are randomly picked to generate N root nodes of initial decision trees. Next, if the node of the decision tree satisfy the splitting condition, some properties in the sample are selected according to several rules. Finally, let all nodes split by the above split attributes until there is no nodes split any more. Through repeating the above steps, multiple abundant decision trees are established to generate a random forest. To calculate their corresponding importance of different decision trees, an importance calculation method of a feature X in the random forest is proposed and described as follows:

First, in the random forest algorithm, each decision tree uses out of bag data (An error estimation method that replace the test set) to calculate the corresponding out of bag data error E_1 . After that, the random forest randomly adds noise interference to the characteristics of all samples of the data out of the bag, which is used to randomly change the value of samples at feature X . Next, the error of the data out of the bag E_2 is calculated. Finally, assuming that there are N decision trees in the random forest algorithm, the importance of feature X is expressed as follows:

$$f(x) = \frac{1}{N} \sum (E_2 - E_1) \quad (1)$$

According to the importance of feature X , data groups with higher relevance and importance are required for maximizing the predicted results. Meanwhile, data groups are also encouraged to maintain less data volume for increasing computational efficiency. The general steps of feature selection are as follows:

First, the characteristic variables in the random forest are sorted in descending or ascending order based on their importance. Next, several extremely low correlation variables in the random forest are eliminated to obtain a new data set. Finally, a new stochastic forest model is established with the new feature group, where the feature importance is calculated as above. Repeat the above steps until M features are left, where M is a pre-defined constant.

3 Experiment and Result Discussion

In this experiment, the terminal device was worn on the experimenter chest as shown in Fig. 2. The LoRa shield wirelessly transmitted human posture data made by the experimenter from the sensor to the Gateway. The human posture data is wirelessly sent to the cloud server using the Gateway through TCP/IP protocol. The human posture data sets were trained on the cloud server using the proposed random forest algorithm. The classification results were sent to the visualization platform through TCP/IP protocol to achieve real-time human posture visualizing.

The cloud server used the random forest method to find out the feature importance of the data. As can be seen in Fig. 3a, the overall importance of acceleration in x axis (1Ax), angular velocity in x axis (2Gx), z axis (2Gz) and magnetometer in z axis (3Cz) exceeded 60%. Therefore, the data in these four axes were selected out to generate a new set of independent variables. As shown in Fig. 3b, c, compared with the raw data performance indicators, the data performance after the first time reduction was significantly improved. Besides, the accuracy of the first time reduction was 87.5%, which slightly larger than that of original data 86.9%. This was proved that the feature selection in the first time reduction was feasible. In contrast, as shown in Fig. 3d, the data performance after the second time reduction declines sharply, which caused the model performance declines very seriously. However, the accuracy of the second time reduction was not reduced much still 86.31%. It can be seen that the second time reduction also was feasible especially under the condition in keeping the accuracy constant.



Fig. 2 Terminal device carried on experimenter to detect real-time posture

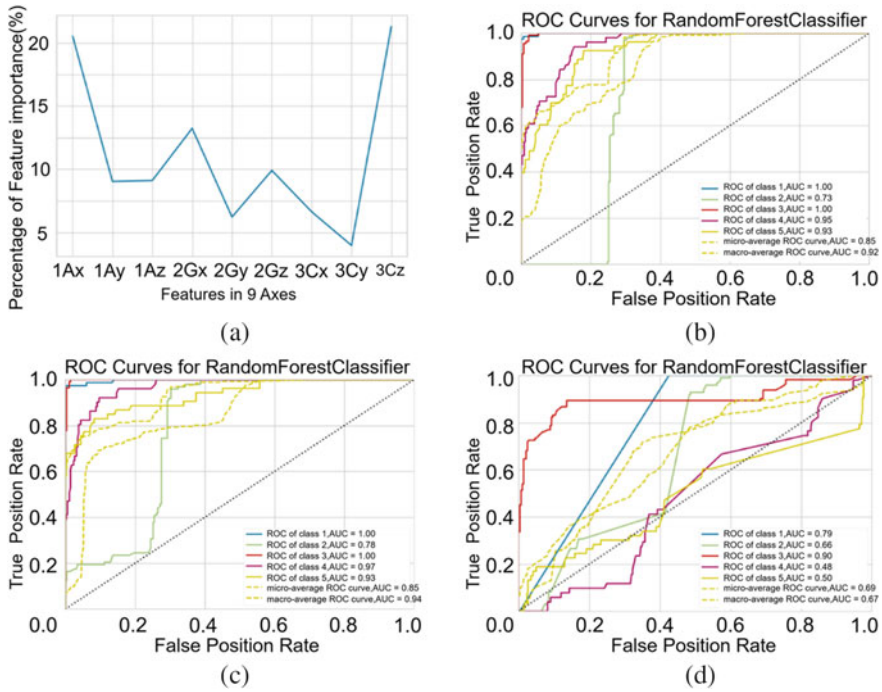


Fig. 3 Data reduction and feature selection. **a** Percentage of each feature importance, **b** raw data performance, **c** data performance after the first time selection, **d** data performance after the second time selection

The visualization system visualized six kinds of human postures on Unity3D using the classification results that trained in the cloud server. First, after the front-end interface of the visualization system was initialized, the program requested the classification results from the cloud server. After that, according to different classification results, Unity3D rendered and updated different animation of human posture, including running, walking, standing, going upstairs, jumping and going upstairs, as shown in Fig. 4. Finally, the visualization system requested the classification results from the cloud server again to display the on-line and real-time posture recognizing results.

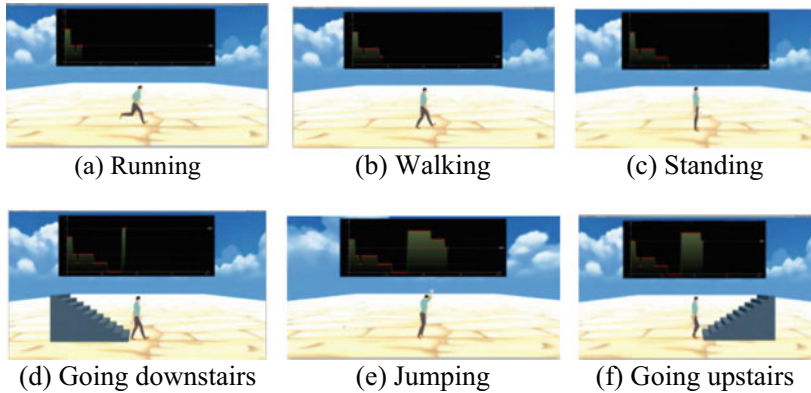


Fig. 4 Visualization of six kinds of human postures

4 Conclusion

In order to realize real-time recognition of human posture, a recognition system based on LoRa protocol is proposed in this paper. Results show that the system accurately recognize six kinds of human postures real-time, which accuracy rate more than 86%. Besides, this paper propose an method of data reduction, which simplify human posture data sets while ensuring accuracy rate. However, the data volume and data type of this experiment is small, so that the generalization ability of the classifier is limited. In the future, the algorithm will be optimized to increase algorithm speed and improve accuracy. Moreover, different data sets of people with different characteristics are collected in quantity to improve the generalization ability of the classifier.

Acknowledgements This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program (2019-0-01585) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), by National Natural Science Foundation of China (61503005), the Great Wall Scholar Program (CIT&TCD20190304, CIT&TCD20190305), Beijing Young Topnotch Talents Cultivation Program (No. CIT&TCD201904009), North China University of Technology “The Belt and Road” Countries Talent Training Base Project, and “Yuyou” Project of North China University of Technology.

References

1. Bao N, Jiang BW, Li YZ (2017) Research on fall detection based on Kinect sensor. *Electron Des Eng* 2017(12):149–152+456
2. Weng PC, Zhang YH (2017) Design of fall detection and positioning system based on MEMS. *Electron Technol Softw Eng* 12:85–86

3. Liu CY, Xu JS, Cheng HT (2015) Attitude detection and data fusion of MPU9250 sensor. *J Henan Univ Sci Technol (Nat Sci Edn)* 2015(04):14–17+22+5
4. Chen XS (2019) Design and implementation of LoRa low-cost full-duplex gateway. *Radio Commun Technol* 02:206–209
5. Ou HJ (2019) A review of machine learning algorithms in the context of big data. *China Inf* 04:50–51

Mobile Charger Planning for Wireless Rechargeable Sensor Network Based on Ant Colony Optimization



Fan-Hsun Tseng, Hsin-Hung Cho, and Chin-Feng Lai

Abstract In order to provide a more flexible wireless rechargeable sensor network, a charger and a self-propelled vehicle are integrated into one vehicle in recent years. The path selection problem of mobile chargers can be formulated as the well-known travelling salesman problem. Therefore, metaheuristic algorithms can be applied to solve the planning problem of mobile chargers. Some researches presented planning methods based on the Simulated Annealing (SA) and Tabu Search (TS) algorithms but the results are not satisfied. In this paper, we not only design a novel encoding approach but also the fitness function for proposing an efficient planning algorithm based on the Ant Colony Optimization (ACO) algorithm. Simulation results show that the proposed ACO-based algorithm achieves a shorter planning path for a longer network lifetime compared with that generated by the SA and TS algorithms.

Keywords Wireless rechargeable sensor network · Wireless sensor network · Power consumption · Network planning

1 Introduction

Wireless rechargeable sensor network (WRSN) enhance the efficiency and enrich the flexibility of wireless sensing network (WSN) [1–3]. For the better charging performance, mobile chargers should be considered while planning wireless chargers

F.-H. Tseng (✉)

Department of Technology Application and Human Resource Development, National Taiwan Normal University, Taipei 16010, Taiwan
e-mail: fhtseng@ntnu.edu.tw

H.-H. Cho

Department of Computer Science and Information Engineering, National Ilan University, Yilan 26047, Taiwan
e-mail: hhcho@niu.edu.tw

C.-F. Lai

Department of Engineering Science, National Cheng Kung University, Tainan 70101, Taiwan
e-mail: cinfon@ieec.org

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_53

387

in an indoor environment. The difference between mobile charging and fixed charging is that mobile charging needs a sophisticated charging schedule. Besides, the cost of fixed charging is more expensive than mobile charging, especially for a low-density environment in a large scale area. Without loss of generality, there is a specific distance between two sensor nodes. It needs to deploy more chargers to fill the coverage holes, thereby the deployment cost of charger increases. Some researchers combine a charger with a self-propelled car and plan an adaptive charging path to guarantee there is no sensor runs out of battery power. However, there are still many research issues need to be solved such as the power reception of sensors, the transmission distance between a sensor node and a charger, obstacles in the terrain, charging priority of sensors.

To charge fixed sensor nodes by a mobile charger is very complex. A simulated-annealing-based method was proposed by other researchers but lack of charging efficiency. In this paper, we utilize a swarm-based algorithm to overcome the drawback of single-solution-based methods. The planning method is proposed on the basis of Ant Colony Optimization (ACO) algorithm. It achieves the better charging efficiency by conducting a shorter charging path for a longer network lifetime than others.

The rest of the paper is organized as follows. Section 2 introduces background and related works. The mobile charging problem is defined in Sect. 3. In Sect. 4, we design an encoding approach and a corresponding fitness function for the proposed Ant Colony Optimization algorithm to solve the defined problem. Section 5 shows simulation results and Sect. 6 concludes this work.

2 Related Works

There are four common wireless charging technologies, i.e., magnetic induction [4], magnetic resonance [5], laser light [6], and microwave conversion [7]. Each of them has pros and cons, and the timing of use. While using the magnetic induction technique, the difference between sensor and charger cannot exceed 5 mm. A general case of charging mobile phones by a wireless charging technique is magnetic resonance. The maximum distance of magnetic resonance is about 3 m long but it is hard to achieve. The laser light technique provides the longer charging distance by using an enormous transmission power, which is more dangerous. The charging distance of microwave conversion technique is higher than 10 m and less influence with environment.

When the distance between sensor nodes is far but charger's charging range is finite, more and more chargers are needed to fill the coverage holes. To solve this problem, the authors in [8] investigated dynamic charger planning in an indoor WRSN environment, which is similar to this work. Chien et al. also try to combine a self-propelled vehicle and a charger into a charging vehicle, thereby the charger vehicle is able to serve more sensor nodes along with the charging path planned. The mobile charging problem can be mapped to a traveling salesman problem (TSP)

which is a well-known NP-complete problem. They utilized the concept of Simulated Annealing (SA) algorithm to plan a charging path. Compared with other meta-heuristic algorithms, the charging method they proposed not only avoids falling into local optimum solution but also finds the optimal path faster. Simulation-based results showed that their proposed SA-based method can make WSN sustainable. In addition, it reduces the cost of charger planning.

However, SA algorithm is a single-solution-based method that still suffers from local optimum phenomenon when the number of sensor nodes increases. Unlike the SA-based approach in [8], we utilized a swarm-based method to overcome the local optimum problem in this paper. We expect that the proposed ACO-based algorithm is superior to the SA-based method in terms of optimizing the path of mobile charger with shorter path and higher efficiency.

3 Problem Definition

The designed charging platform in this paper is captured in Fig. 1. Sensor nodes are represented by red points and are distributed in the indoor space randomly. The platform is an obstacle between the charging vehicle and sensor nodes. Note that the sensor can be charged normally as long as the platform is a nonmetal platform, which has lower impact on charging quality. The received power of the i -th sensor can be calculated by

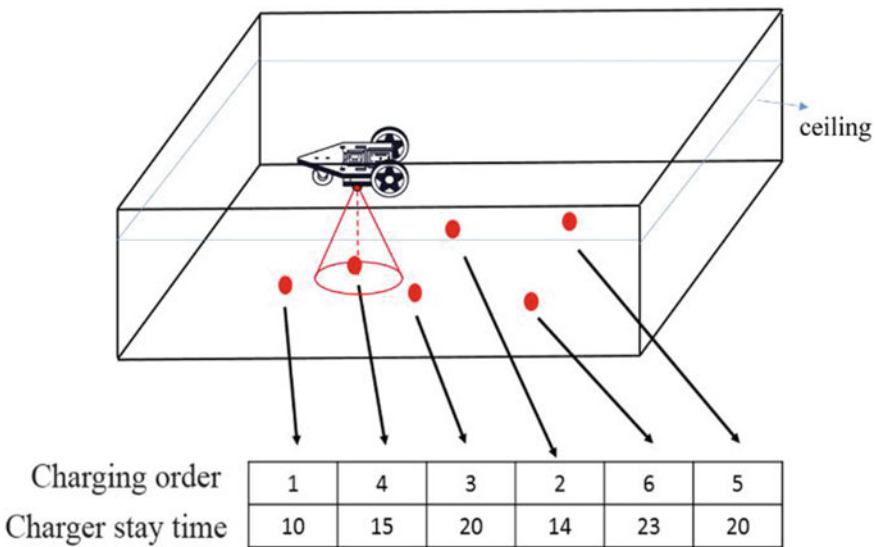


Fig. 1 Schematic diagram of WRSN charging planning

$$P_i^s = \frac{G_c G_s \eta \sigma}{L_p} \left(\frac{\lambda}{4(d_i + \beta)} \right)^2 P^c, \quad (1)$$

where the variable σ correlates with the used platform material. The distance between the charging vehicle and the sensor is represented by d_i . The notations G_c and G_s represent the antenna gains of charger and sensor. The notation η stands for rectifier efficiency and the notation L_p represents polarization loss. The notation λ is the wavelength of RF and the notation β is an adjusted parameter in an indoor scenario. The transmission power from a charging vehicle to a sensor is represented by P^c .

A Boolean value S_i is defined to present the status of a sensor node. When $S_i=1$ means that the charging vehicle cannot meet the i -th sensor's request thereby the i -th sensor will be dead, and vice versa. The Boolean value S_i is defined as

$$S_i = \begin{cases} 1, & B_i + P_i^s \times t_i^c < w_i^s \times t_i^u \\ 0, & B_i + P_i^s \times t_i^c \geq w_i^s \times t_i^u \end{cases}, \quad (2)$$

where B_i is the remaining power of the i -th sensor node. The notation t_i^c represents the required time for charging of the i -th sensor node. Note that a sensor node still consumes its power while a charging vehicle is charging for itself. The notation t_i^u represents the time period of the i -th sensor node during the charging vehicle approaches and leaves the i -th sensor node. The notation w_i^s is the power consumption rate of the i -th sensor node. Each sensor node has its power consumption rate. To exactly design a charger planning, we should calculate the required charging time t_i^c of each sensor node. While $S_i = 1$, the required charging time of the i -th sensor node as well as the notation t_i^c can be derived by

$$t_i^c = \frac{t_i^u \times w_i^s - B_i}{P_i^s}. \quad (3)$$

The traveling time of a charging path is also a vital issue to the path planning of WRSN. It is denoted as t_{travel} , which can be calculated by

$$t_{travel} = \sum_{j=1}^M t_j^r + \sum_{i=1}^N t_i^c, \quad (4)$$

where the notation t_j^r represents the time of the charging vehicle for driving along with the j -th path. The notation N is the total number of sensor nodes. Let $C = \{c_1, c_2, \dots, c_N\}$ be a set of sensor nodes, and c_i records the coordinate of the i -th sensor nodes. Let $V = \{v_1, v_2, \dots, v_N\}$ be a set of charging schedule, and v_1 represents the first sensor node which will be charged. Let $T = \{t_1, t_2, \dots, t_N\}$ be a set of charging time for sensor nodes.

The planning problem of WRSN in this work is formulated based on linear programming model, which is formulated as

$$\min \sum_{i=1}^N S_i$$

s.t.

$$t_i^u \geq t_{travel},$$

$$d_i < R.$$

The major goal of this study is to minimize the number of dead sensor nodes. Some restrictions are described and explained as follows. The equation $t_i^u \geq t_{travel}$ means that the maximum used time of each sensor must be longer than the time of a charging travel. It guarantees that each sensor node will not die while the charging travel is still in progress. The equation $d_i < R$ guarantees that the charging distance between a mobile charger and a sensor node is an effective charging distance.

4 Proposed Scheme

In this paper, a swarm-based algorithm is proposed to solve the defined planning problem of WRSN. Firstly, a table is created to record all the positions passed by the ants. The value is denoted as 0 once an ant visits the position, otherwise, its value will be changed to 1. The first charged sensor has the highest charging demand. Then, the charging path will follow the roulette wheel selection. It will be changed when the fitness function changes. Once an ant selects a path, the corresponding positions which have been walked by other ants in the check list will be set to 0. In this way, duplicate paths can be avoided. When a path has been determined and ants start to act, the pheromone will be updated. If there is a path causing the node to die, the pheromone of such node will be increased. In the proposed ACO-based algorithm, the fitness function depends on pheromone calculation, which is calculated by

$$\frac{P^\alpha \times \frac{1}{d}^\beta \times N_d}{\sum_{i=1}^n P^\alpha \times \frac{1}{d}^\beta \times N_d}, \tag{5}$$

where P^α represents the value of pheromone and $\frac{1}{d}$ is the reciprocal of the distance. It implies that a farther distance has less influence. The details are captured in Fig. 2.

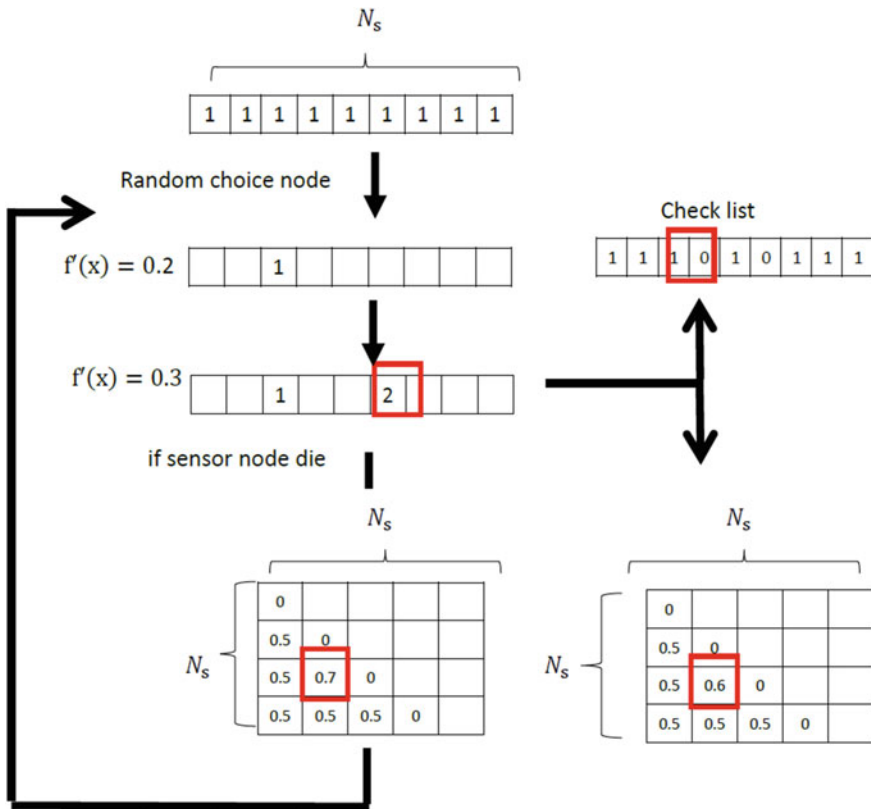


Fig. 2 Proposed ACO-based charging method

5 Simulation Results

The results are simulated and performed by using MATLAB (Version 7.11, R2010b). There are 20–120 sensor nodes in a 20 × 25 m square outdoor environment. The speed of charging vehicle sets to 2 m per unit time. The consumption power of each sensor node was set 0.01–0.02 W per unit time. The charging efficiency of charging vehicle sets to 2 W per unit time.

The result of path length is captured in Fig. 3. It can be observed that the proposed ACO-based algorithm yields the shortest charging path compared with that generated when using SA and TS algorithms. This is attributed to the fact that the proposed ACO-based algorithm has a TSP operator. The result of number of dead sensors is captured in Fig. 4. It is a main indicator for evaluating whether these methods have the ability to solve the charging problem. More dead sensor nodes imply that the network lifetime of WRSN is shorter. The proposed ACO-based algorithm is capable of handling the mobile charging problem. It yields more remaining sensor

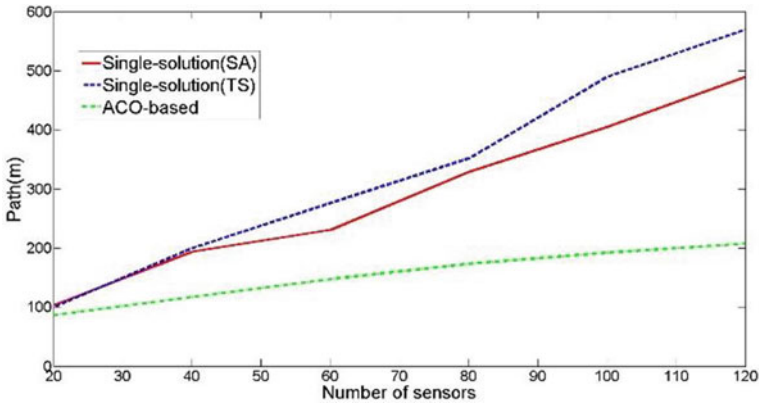


Fig. 3 Results of path length

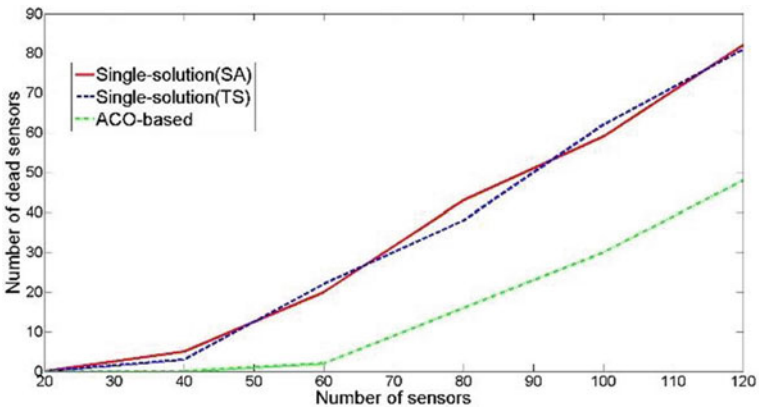


Fig. 4 Results of the number of dead sensors

nodes compared to the SA and TS algorithms. This is attributed to the fact that the SA cannot reflect with the restrictions of continuity.

6 Conclusion

Energy consumption is a vital issue to wireless rechargeable sensor networks but fixed chargers always serve few sensors which is inefficient to cost and performance. A more flexible charging method as known as the mobile charger deployment has been proposed. In this paper, we designed that an automatic vehicle is able walk around the ceiling mezzanine. The automatic vehicle equips a charger to electrify sensor nodes. As a result, the fixed charger placement problem turns into a TSP. Since it has

been proved that TSP is an NP-hard problem, any single-solution-based algorithm cannot provide the best charging route. The paper presents a new encoding approach for the proposed ACO-based algorithm, and a corresponding fitness function for evaluating the charging route. Simulation results showed that the proposed ACO-based algorithm is superior to other algorithms in terms of charging path length and network lifetime. In the future, we expect to propose other meta-heuristic algorithms to optimize the charging path of wireless rechargeable sensor networks.

Acknowledgements This work was financially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant MOST108-2636-E-003-001, and was partly funded by the MOST in Taiwan, under grant MOST108-2221-E-197-012-MY3.

References

1. Habib C, Makhoul A, Darazi R, Salim C (2016) Self-adaptive data collection and fusion for health monitoring based on body sensor networks. *IEEE Trans Ind Inf* 12(6):2342–2352
2. Tseng FH, Cho HH, Chou LD, Chao HC (2013) Efficient power conservation mechanism in spline function defined WSN terrain. *IEEE Sens J* 14(3):853–864
3. Cho HH, Shih TK, Chao HC (2015) A robust coverage scheme for UWSNs using the spline function. *IEEE Sens J* 16(11):3995–4002
4. Stielau OH, Covic GA (2000) Design of loosely coupled inductive power transfer systems. In: *Proceedings of the international conference on power system technology*, pp 85–90
5. Kiani M, Ghovanloo M (2012) The circuit theory behind coupled-mode magnetic resonance-based wireless power transmission. *IEEE Trans Circuits Syst I Regul Pap* 59(9):2065–2074
6. Ching NN, Wong HY, Li WJ, Leong PH, Wen Z (2002) A laser-micromachined multi-modal resonating power transducer for wireless sensing systems. *Sens Actuators A* 97:685–690
7. Shinohara N (2011) Power without wires. *IEEE Microw Mag* 12(7):S64–S73
8. Chien WC, Cho HH, Lai CF, Shih TK, Chao HC (2017) Dynamic charging planning for indoor WRSN environment by using self-propelled vehicle. In: *Proceedings of the international conference on knowledge management in organizations*, pp 547–559

Deep Learning Based Malware Analysis



Sunoh Choi

Abstract Today, hundreds of thousands of new malicious files are being made. The existing pattern-based antivirus solution has difficulties in coping with such a large number of new malicious files. To solve these problems, artificial intelligence based malicious file detection methods have been proposed. In this paper, we propose a malicious file analysis method based on deep learning.

Keywords Malware analysis · Deep learning · Attention

1 Introduction

Nowadays hundreds of thousands of malicious files are being created every day [1]. Also, malicious files using zero-day vulnerabilities are being created [2]. For this reason, existing pattern-based antivirus solutions have difficulty responding to new malicious files [3].

Traditional pattern-based antivirus solutions determine malicious files based on hash values of malicious files, special string that appear in malicious files, or malicious file behavior. However, new malicious files are designed to avoid detection of existing antivirus solutions [4].

Recently, methods for detecting malicious files using artificial intelligence have been studied [3–9]. Artificial intelligence has been widely used for image recognition and machine translation [10, 11]. Machine learning and deep learning techniques are widely used in artificial intelligence. The advantage of artificial intelligence is that you can make decisions about similar data through learning. For example, if feature data of a new file similar to the feature data of a previously learned malicious files comes in, the file can be judged as a malicious file by artificial intelligence. That is, even if there is no pattern for the new malicious file, it is possible to judge whether or not the new malicious file is malicious.

S. Choi (✉)

Honam University, 8510 Changjo-gwan, Gwangju, South Korea

e-mail: suno@honam.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_54

395

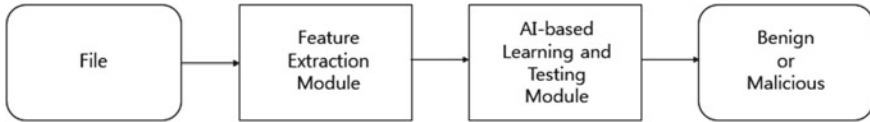


Fig. 1 AI-based malware detection system structure

In order to detect malicious files based on artificial intelligence, two modules are required in Fig. 1. First is a malicious file feature extraction module and second is artificial intelligence based learning and testing module. The malicious file feature extraction module is a module for extracting feature data from malicious files. AI based learning and test module is a module for learning and testing AI based model with extracted data. Although it is important to make artificial intelligence-based learning and testing modules well in artificial intelligence based malicious file detection systems, extracting features from malicious files is also very important for improving system accuracy.

In this paper, we propose a malicious file feature extraction method based on deep learning [11]. The deep learning based method calculates the weight of each input value of the deep learning model to determine which input value has more influence on the result value.

2 Deep Learning Based Malware Detection

2.1 Feature Extraction Using Static Analysis

We need malicious file feature data for malicious file detection using deep learning. For this, we can extract malicious file feature data using static analysis. For static analysis, we extract the assembly code as shown in Fig. 2 using objdump [12]. Next, we extract opcode sequences such as push, mov, and sub. We can then construct a

```
401000: 55                push   %ebp
401001: 8b ec             mov    %esp,%ebp
401003: 83 ec 5c          sub   $0x5c,%esp
401006: 83 7d 0c 0f      cmpl  $0xf,0xc(%ebp)
40100a: 74 2b             je     0x401037
40100c: 83 7d 0c 46      cmpl  $0x46,0xc(%ebp)
401010: 8b 45 14          mov   0x14(%ebp),%eax
401013: 75 0d             jne   0x401022
401015: 83 48 18 10      orl   $0x10,0x18(%eax)
401019: 8b 0d a8 3e 42 00 mov   0x423ea8,%ecx
```

Fig. 2 PE file assembly code

trigram sequence [13] for three consecutive opcodes. The trigram sequence is created as follows:

(push, mov, sub), (mov, sub, cmpl), (sub, cmpl, je), ...

The reason for creating trigram sequences is that there are approximately 100 opcodes, and when these are placed into trigrams, the size of the trigram domain is approximately 100^3 , such that the trigram sequence of each file can easily be distinguished from those of other files.

2.2 Attention

Attention is a deep learning mechanism that looks for the parts of sequence data having greater impacts on the results. A typical example of attention is text summarization [14] which involves summarizing a given text. Using the attention mechanism, we can identify some of the main words to summarize an article when it is given as a sequence of words. For example, the text shown in Fig. 3 is summarized as follows:

Russia calls for joint front against terrorism

The RNN model is utilized in neural machine translation (NMT). For example, German sentences can be translated into English. NMT encodes the source sentence into a vector, and decodes the sentence based on that vector.

The attention mechanism allows the decoder to refer to a portion of the source sentence as shown in Fig. 4 [15]. Here, X is the source sentence and y is the translated sentence generated by the decoder. Figure 4 depicts a bidirectional recurrent neural network. The important point is that the output word y_t depends on the weight combination of all input states. Here, α is a weight that defines how strongly each

Fig. 3 Alignment of text summarization [14]

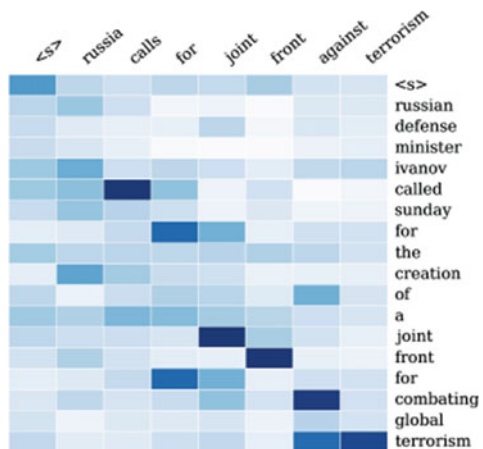
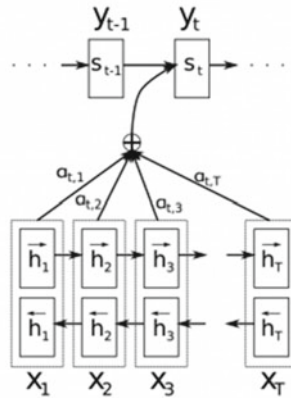


Fig. 4 Attention [15]



input state influences each output state. If $\alpha(3, 2)$ is a large value, then this means that the third word in the output sentence refers to the second word in the input sentence.

The advantage of attention is that it provides the ability to interpret and visualize what the model does. For example, by visualizing the attention weight matrix when a sentence is translated, we can understand how the model performs translation. As shown in the text summarization above, one can observe how strongly each word in the output summary statement refers to each word in the input sentence.

3 Deep Learning Based Malware Analysis

This section proposes a deep learning based malware analysis method. The idea behind this method is as follows. When we utilize the attention model to identify the weight of each API system call in a sequence of length n , we consider subsequences of length k by extracting weighted words for malicious file detection.

For example, the API sequence for a malicious file might look as follows:

{LoadLibrary, LoadCursor, RegisterClass, GetThreadLocal, strcmp, GlobalAlloc, GlobalFree, FindResource, LoadResource, VirtualProtect}

In the attention model, the weight of the sequence data may be given as follows:

{0.3, 0.0125, 0.0125, 0.0125, 0.3, 0.0125, 0.0125, 0.0125, 0.0125, 0.3}

Here, we extract the APIs with the top-3 weights:

{LoadLibrary, strcmp, VirtualProtect}

This is the API sequence pattern that appears in import address table hooking malicious files [5]. Then, after extracting the important data subsequences, malicious file analysts can more easily analyze malicious files. In other words, malicious file analysts can analyze malicious files by analyzing important subsequences of data sequences rather than having to examine the entire data sequence.

4 Conclusion

In this study, we propose a deep learning based malware analysis method. The attention based malware analysis method allows malicious code analysts to only analyze parts of malicious code based on the features extracted by the attention mechanism rather than analyzing the entire malicious code. This is expected to considerably reduce the efforts required by malicious code analysts.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1G1A11100261).

References

1. AV-TEST. <https://www.av-test.org>
2. Zero-day. https://en.wikipedia.org/wiki/Zero-dat_computing
3. Gavrilut D, Cimpoesu M, Anton D, Ciortuz L (2009) Malware detection using machine learning. In: International multicongress on computer science and information technology
4. Gibert D (2016) Convolutional neural networks for malware classification. Master thesis, Universitat de Barcelona
5. Ki Y, Kim E, Kim HK (2015) A novel approach to detect malware based on API call sequence analysis. *Int J Distrib Sens Netw*
6. Saxe J, Berlin K (2015) Deep neural network based malware detection using two dimensional binary program features. In: International conference on malicious and unwanted software (MALWARE)
7. Dahl GE, Stokes JW, Deng L, Yu D (2013) Large-scale malware classification using random projections and neural networks. In: International conference on acoustics, speech and signal processing (ICASSP)
8. Pascanu R, Stokes JW, Sanossian H, Marinescu M, Thomas A (2015) Malware classification with recurrent networks. In: International conference on acoustics, speech and signal processing (ICASSP)
9. Huang W, Stokes JW (2016) MtNet: a multi-task neural networks for dynamic malware classification. In: International conference on detection of intrusions and malware and vulnerability assessment (DIMVA)
10. Kaiming H, Ziangyu Z, Shaoqing R, Jian S (2016) Deep residual learning for image recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)
11. Wang Y, Tian F (2016) Recurrent residual learning for sequence classification. In: International conference on empirical methods in natural language processing (EMNLP)
12. Objdump. <https://en.wikipedia.org/wiki/Objdump>
13. n-gram. <https://en.wikipedia.org/wiki/N-gram>

14. Rush AM, Chopra S, Weston J (2015) A neural attention model for sentence summarization. In: International conference on empirical methods in natural language processing
15. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 6th international conference on learning representations

CNN-GRU-Based Feature Extraction Model of Multivariate Time-Series Data for Regional Clustering



Jinah Kim and Namme Moon

Abstract Clustering-related research on data with time continuity is largely done through statistical analysis and thus does not fully reflect the data's features. In this paper, we propose a CNN-GRU-based model to extract each variable's time-dependent changes and features in multivariate data. We have utilized CNN to identify the features of each variable and derive trends over time based on GRU. Fuzzy C-means clustering is performed based on this feature and overlapped cluster results are finally obtained. Experiments were conducted using two years of card usage data to extract the features according to the local consumption industries and apply these to regional clustering. The proposed method's performance is evaluated by comparing the proposed method with data characterization and clustering methods used in existing research.

Keywords Multivariate time-series clustering · Regional clustering · CNN · GRU · Fuzzy C-means clustering

1 Introduction

Research on time-series data mining has been performed in various fields such as prediction, pattern search, rule discovery, classification, and clustering, and has been applied in many fields such as weather, stocks, and medical care [1, 2]. Since time-series data have irregular fluctuations according to their trends, they have nonlinear features. To take this into consideration, researchers are changing from focusing on statistical models such as Auto-regressive Integrated Moving Average (ARIMA) to neural networks [3, 4].

J. Kim

Department of Computer Engineering, Hoseo University, Asan-si 31499, South Korea
e-mail: jina9406@gmail.com

N. Moon (✉)

Division of Computer Information Engineering, Hoseo University, Asan-si 31499, South Korea
e-mail: nammee.moon@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_55

401

Meanwhile, research into clustering in time-series data mining can be broadly divided into data with similar patterns and research that clusters time; it is necessary to grasp the features of time-series data over time. In recent years, Convolutional Neural Network (CNN), which is a type of deep network, has been widely employed to extract the features of time-series data. CNN is known to show good performance for automatically extracting features for input data and is widely used in image data. In particular, it is suitable for extracting the features of time-series data because it is robust against data deformation [5].

Thus, in this paper, we want to understand the features of multivariate time-series data based on CNN. In this instance, we propose a feature extraction method that can reflect the trends of the entire time series by using the Recurrent Neural Network (RNN)-based Gated Recurrent Unit (GRU) model that specializes in time-series data. GRU has the advantage of faster learning and a simpler structure while improving the gradient vanishing problem of RNN compared to other RNN-based methods [6]. Finally, based on the extracted features, we aim to achieve effective clustering that can reflect the multivariate trends in the same time period. In this study, we applied the proposed method to cluster regions with similar consumption patterns by utilizing regional card data from 2017–2018.

2 Regional Clustering Through Feature Extraction from Multivariate Time-Series Data

This study aims to extract and type features from multivariate time series data and to cluster regions with similar features according to time series. In this case, it is different from previous studies in that a new feature sequence is created by extracting a feature for each specific time interval using multivariate time series data.

In this study, by extracting the features of consumption trends according to the regional consumption category, regions with similar consumption amounts and trends are clustered. There are eight categories of consumption: restaurant, life, sports/culture/leisure, travel/transportation, education, medical, home appliances/furniture, and automotive. This differs from the conventional clustering method because it sequences trends over time and consumption amounts by category.

Figure 1 shows the proposed clustering method based on feature extraction from multivariate time-series data; data preprocessing, feature extraction, and clustering are performed.

First, preprocessing removes the noise of the data to grasp the trend. Smoothing is performed using Moving Average (MA), which is the most basic method to show a trend by eliminating the noise included in the data. The average of m observations is calculated based on period m . The larger m is, the more the period disappears. Therefore, choosing an appropriate size for m is important. Since the data used in this study is day data, we set seven days a week. Since the amount of consumption by category is large, the scale was from 0 to 1, and the subsequence was then extracted

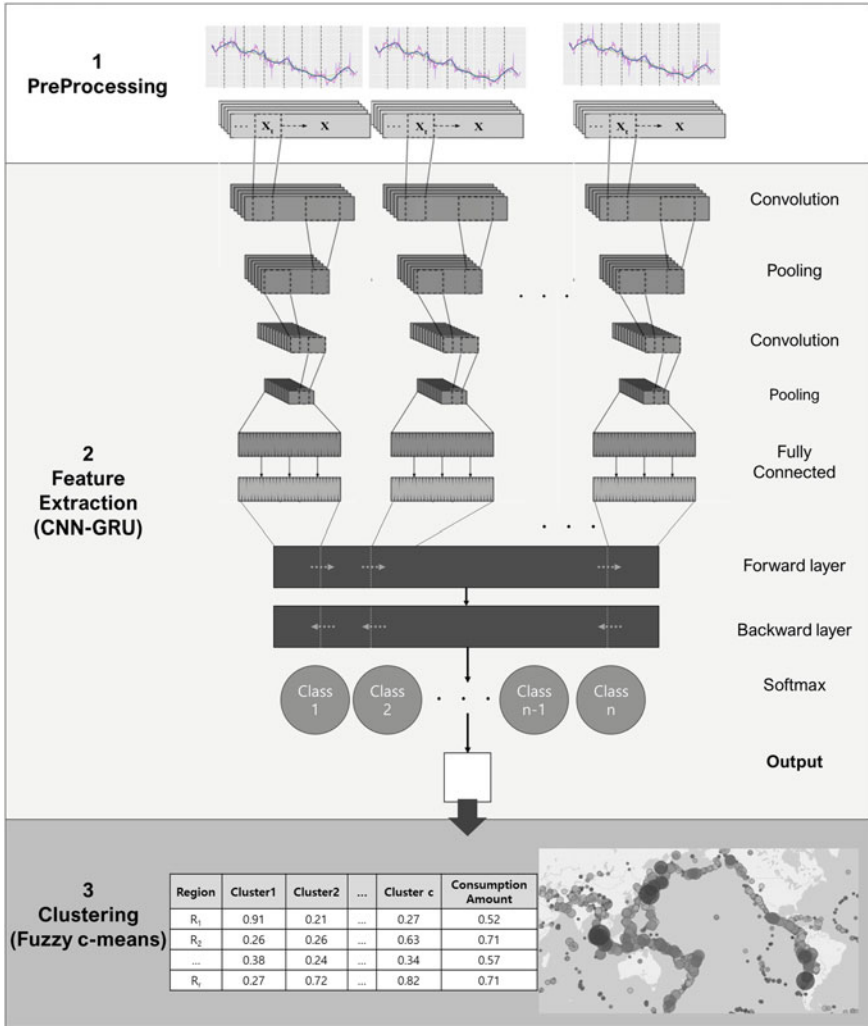


Fig. 1 Process of clustering through feature extraction from multivariate time-series data

from the entire time series by setting a specific data point at a constant time interval for feature extraction. Each subsequence is input into the neural network for feature extraction.

Next, CNN and GRU are combined for each subsequence to extract features for consumption trends by region. This process works to identify consumption trends by category rather than consumption amount. Section 3 gives a description of the neural network structure.

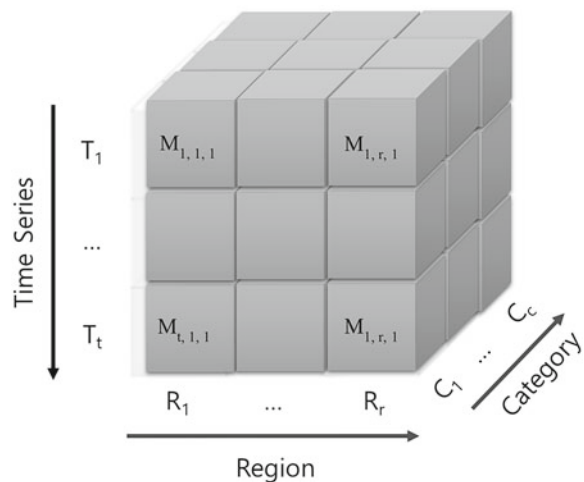
When the feature extraction was completed, each region had the result of extracting the consumption trend feature according to the type of industry. Similarity was calculated for each category and overlapping clustering based on Fuzzy C-means was performed between similar regions. The reason for overlapping clustering is that there exists a possibility that the trends of various categories by region may have tendencies of various clusters.

3 Feature Extraction of Time-Series Data Based on CNN-GRU

To extract a feature, CNN is first performed to extract the features of the subsequence according to the category and the trend pattern is classified through the GRU. The reason for using GRU is that it is simpler than the commonly used Long Short-Term Memory models (LSTM) models for dealing with time series data but is known to have similar performance [7]. In this study, the GRU model is more suitable because it deals with multivariate time series data.

The input data consists of $T \times R \times C$ three-dimensional neurons as shown in Fig. 2. T is the length of the time series, R is the region, and C is the consumption category. Next, the features are detected through the convolutional layer and the feature map is output. We use the Max Pooling layer to reduce the size of the feature data and repeat this process several times. Finally, the data size and dimensions are lowered by flattening the layers and both the input and output are connected by the dense layer. Thus, the features of each subsequence are linked and sequenced. This result is reflected as an input value of GRU learning and the trend flow for each subsequence can be grasped. The GRU learning model computes both forward and

Fig. 2 Input data structure



backward using two hidden layers to utilize both previous information and future information on the current time basis; its type is many to one. Finally, it receives the values from two hidden layers in the output layer and classifies the trend of the corresponding category.

4 Conclusion

In this paper, we have proposed a method for extracting the trend features of multivariate time-series data based on CNN-GRU for clustering observations with multivariate time-series data. Data was used by the card in accordance with the category by region and applied to the clustering of regions that show similar consumption trends. This differs from previous studies in that it is patterned to reflect the whole flow of multivariate time series.

The card data used in this study has the feature that it fluctuates due to social influence. However, there is a limit in that it does not add variables that can take into account change in trends over time. Future research will be able to reflect social influences using data such as SNS or news and confirm the degree of impact in each region. In addition, we want to expand the regional feature to the recommendation system in offline consumption through clustered results.

Acknowledgements This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2019 (R2018020083).

References

1. Fu TC (2011) A review on time series data mining. *Eng Appl Artif Intell* 24(1):164–181
2. An HW, Moon N (2019) Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM. *J Ambient Intell Hum Comput* 1–11
3. Li L, Wu Y, Ou Y, Li Q, Zhou Y, Chen D (2017) Research on machine learning algorithms and feature extraction for time series. In: 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), pp 1–5
4. Büyüksahin ÜÇ, Ertekin Ş (2019) Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. *Neurocomputing* 361:151–163
5. Zhao B, Lu H, Chen S, Liu J, Wu D (2017) Convolutional neural networks for time series classification. *J Syst Eng Electron* 28(1):162–169
6. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. [arXiv:1803.01271](https://arxiv.org/abs/1803.01271)
7. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. [arXiv:1412.3555](https://arxiv.org/abs/1412.3555)

DNN-Based Mutual Satisfaction Prediction Model for Matching Between Users



Hyunnoh Yun, Jinah Kim, and Nammee Moon

Abstract Recently, there have been many attempts to provide customized services using big data. In this paper, we try to predict the degree of mutual satisfaction for matching between users. First, the proposed system calculates the supplier's expected satisfaction and then calculates the higher-result supplier's expected satisfaction. The participants' personal information, preference information, and mutual evaluation are utilized as input values for learning. By using mutual evaluation for learning, recommendations can be made that consider both directions rather than unilateral recommendations. The mutual satisfaction level is obtained by suggesting both the consumer's and supplier's expected satisfaction. This results in better satisfaction scores for users and their matching systems.

Keywords Deep learning · DNN · Mutual satisfaction · Matching between users

1 Introduction

Providing users with personalized services and recommendations is an important aspect of a company's marketing. Therefore, attempts and research to provide personalized services using big data are continuously carried out [1–3]. Most studies match users and products to provide recommendation services. User-to-product matching uses methods such as content-based filtering and collaborative filtering [4–6].

However, applying the aforementioned methods to recommendations between users is difficult [7]. First, users have different variables, making it difficult to get

H. Yun · J. Kim

Department of Computer Engineering, Hoseo University, Asan-Si 31499, South Korea
e-mail: zxcv9153@naver.com

J. Kim

e-mail: jina9406@gmail.com

N. Moon (✉)

Division of Computer and Information Engineering, Hoseo University, Asan-Si 31499, South Korea
e-mail: nammee.moon@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_56

407

the same type of data. Second, the supplier’s capriciousness also acts as a variable. A supplier may not make the same choices in requests under the same conditions. Finally, this is a relationship in which mutual evaluation takes place rather than unilateral choices, such as in the relationship between users and products. Therefore, it is possible to expect a higher level of satisfaction when a mutual recommendation score is reflected simultaneously.

This paper presents a recommendation service model between users based on deep running. Statistical analysis derives what factors affect participants’ choices. The participants’ personal information, preference information, and mutual evaluation scores are used as learning variables by referring to what factors affect their choices. Through the two Deep Neural Network (DNN) models, mutual satisfaction can be obtained by predicting the satisfaction of both the consumer’s and supplier’s positions.

2 System Architecture

Figure 1 shows the system’s proposed structure. First, based on surveys and statistics, the factors affecting the choices between consumers and suppliers are analyzed. Elements are divided into filtering and weighting properties. Property filtering is the process by which participants filter what data meets the desired criteria from the overall data. The weighting properties reflect the weighting factors used in learning

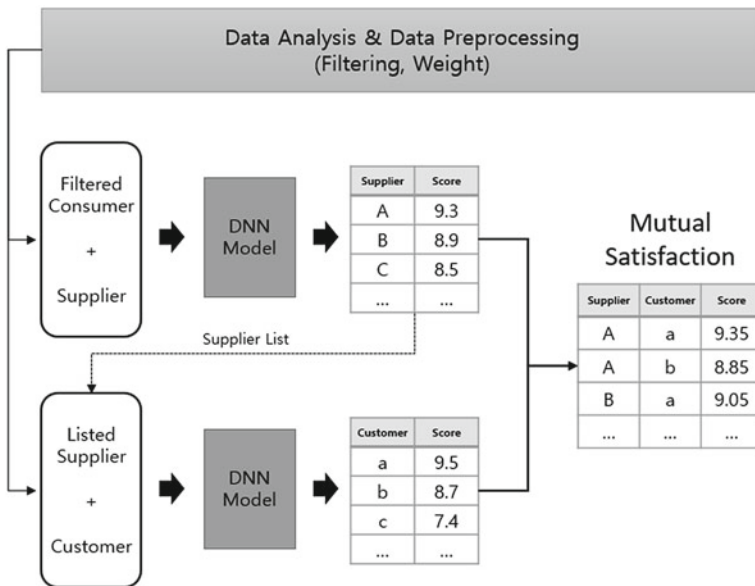


Fig. 1 System architecture

based on data analysis and the weighting characteristics of filtered consumers and suppliers are obtained by driving in-depth with input values. Supplier satisfaction is obtained using a list of the top providers and consumer information from the acquired consumer satisfaction. Mutual satisfaction is achieved using a formula that suggests consumer satisfaction and supplier satisfaction printed in both models.

3 The Proposed Method

3.1 Data Analysis and Data Preprocessing

This experiment was conducted based on the matching system of art education. For the data analysis, the survey of 106 questions on arts education to 800 people and integrated data on education-related card consumption from January 2017 to August 2018. Statistical analysis and cluster analysis were used as data analysis methods. The results of a cluster analysis of survey data by income group and statistical analysis of survey data and card consumption data were reflected in the filtering and weighting properties among survey participants.

In addition, participants' activity data were used; i.e. data generated by the use of participants' personal information and services. In particular, two-way assessment data are used rather than one-sided, thus using evaluation data between consumers and suppliers.

Table 1 shows the filtering and weighting properties to be used for learning that reflect the analysis results.

3.2 Deep Neural Network

A DNN is a structure with two or more layers of concealment in the structure of an existing Artificial Neural Network (ANN) as shown in Fig. 2. ANN uses a model that evaluates the experience rather than direct programmers to set existing rules and a set of commands to correct the model when it makes a mistake. The data entered into the input layer is output to the output layer by modifying the weight value through the activation function in the hidden layer. Unlike ANNs with a single hidden layer, DNNs can be classified or predicted at the output layer using features abstracted from several hidden layers for complex problems.

In this paper, the input values are the attributes given in Table 1. The two DNN models modify the weights of each node in the hidden layer using participant data entered into the input values. Both models predict participants' expected scores and calculate errors to proceed with the learning process. At the end of the learning process, both the users' and suppliers' expected scores are printed out.

Table 1 Filtering and weight properties table

Model	User	Property	Element
Customer matching model for supplier	Supplier	Personal information	Age
			Location
		Preference information	Position difference with the consumer
	Customer	Personal information	Age
Evaluation information		Supplier's score	
Supplier matching model for customer	Customer	Personal information	Gender
			Age
			Location
		Preference information	Lesson price
	Position difference with the Supplier		
	Supplier career		
	Supplier	Personal information	Gender
			Age
Movable distance			
Career			
	Evaluation information	Customer's score	

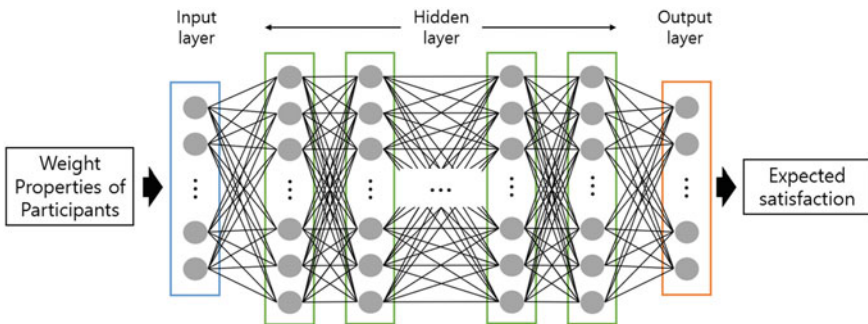


Fig. 2 Deep neural network model architecture

3.3 Mutual Satisfaction

$$MS_{x,y} = \frac{(x + y) - \sqrt{\left(\frac{x+y}{2} - x\right)^2}}{2} \tag{1}$$

Mutual satisfaction refers to satisfaction with both the consumer and supplier; it does not only consider the customer's satisfaction. This is meaningful in reflecting

both the consumer's and supplier's assessment in recommendation matching between users. However, calculating two assessments as averages can cause problems; for example, assume the satisfaction of participants with two or more equal averages. Even with the same mean value, the greater the difference between two values, the more likely that the lower satisfaction of one participant may be ignored, making it difficult to state a good match. Therefore, a formula is needed to obtain a satisfaction score for two participants and reciprocal satisfaction that reflects their differences. In this paper, the mutual satisfaction level is calculated using Eq. (1). The calculation formula is obtained by subtracting the variance from the sum of two numbers and dividing by two. This reflects the difference between the two values using variance, which is the deviation from the average. By calculating the variance from the sum of the two numbers, the larger the difference between the two values, the smaller the value. Dividing them by two will obtain the mean, which will reflect the difference between the two values.

4 Conclusion

Based on the Art Education Matching System, this paper proposes a matching system between users. The DNN model is applied to predict the participants' expected satisfaction to match users, unlike the recommended services for existing users and products. The input value of DNN reflects what factors are derived from the data analysis. A mutual satisfaction level can be obtained that reflects the evaluation of both parties with the satisfaction predicted by the model. Mutual satisfaction, which reflects the two sides' assessment, can expect better satisfaction than one-sided user-user matching. However, if there is insufficient data to be used for learning, the model may not be able to learn. A future plan is to study ways to implement an improved model that applies CNN's feature search.

Acknowledgements This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2019 (R2018020083).

References

1. Lee H, Lee W (2018) A study on the design and implementation of the learned life sports team recommendation service system based on user feedback information. *J Korea Multimed Soc* 21(2):242–249
2. So K, Lee Y, Moon K, Ko K (2015) Design of bi-directional recommend calligraphy contents open-market platform. *J Korea Multimed Soc* 18(12):1586–1593
3. Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anil R (2016) Wide & deep learning for recommender systems. In: *Proceedings of the 1st workshop on deep learning for recommender systems*. ACM, pp 7–10

4. An H, Moon N (2019) Influential factor based hybrid recommendation system with deep neural network-based data supplement. *J Broadcast Eng* 24(3):515–526
5. Zheng L, Noroozi V, Yu PS (2017) Joint deep modeling of users and items using reviews for recommendation. In: Proceedings of the tenth ACM international conference on web search and data mining. ACM, pp 425–434
6. Wang H, Wang N, Yeung DY (2015) Collaborative deep learning for recommender systems. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM pp 1235–1244
7. How Airbnb uses machine learning to detect host preferences. <https://medium.com/airbnb-engineering/how-airbnb-uses-machine-learning-to-detect-host-preferences-18ce07150fa3>

An Approach to Improving Software Security Through Access Control for Data in Programs



Hyun-il Lim

Abstract In a recent information system, effective management of secure information plays an important role in maintaining the system securely. To ensure the safe operation of a system, secure information in the system should be kept safely and prevented from external intrusion or information leakage. So, it is needed to protect secure information against unauthorized access to maintain a safe information system. In this paper, we present a data secure language and design an access control method for protecting secure data against unauthorized access in programs. The proposed method is designed to manage data containing secure information, and it can improve the information security of programs. In experiments, we show evaluation results and the accuracy of the proposed method.

Keywords Data access control · Software security · Software analysis · Information analysis

1 Introduction

Recently, information and communication technologies are being used in a variety of areas, with the rapid advancement of data manipulations in software. In addition, vast amounts of data and information are produced and used in the present society, and they play an important role in operating information systems. On the other hand, efficient and secure manipulation of data is required to improve the security in such systems. For example, if important data is exposed to users without authorization in software execution, there will be security risks in the software system. Thus, a security model that protects important data from unauthorized access is required in software systems to improve system security [1–5].

In this paper, we propose an approach to controlling access to data used in software to protect secure data from unauthorized access. We design a data secure language

H. Lim (✉)

Department of Computer Engineering, Kyungnam University, Gyeongsangnam-do 51767, South Korea

e-mail: hilim@kyungnam.ac.kr

and propose a method for controlling accesses to secure data. We also perform experiments and evaluate the efficacy of the proposed method.

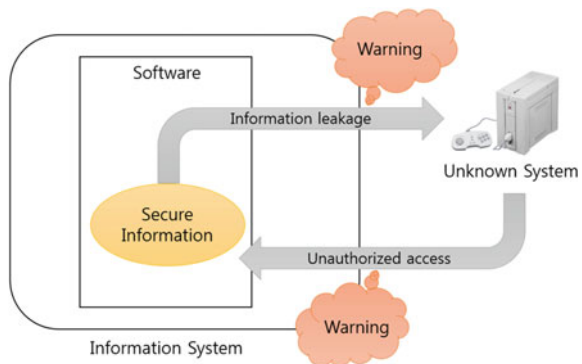
2 Data Access Control

The data that are used for computation in a program are generally passed as variables or through memory. In operating information systems, data have essential information for information service, and the data should be maintained securely so that secure data cannot be accessed without permission. For example, if private information, such as security numbers containing personal information, is leaked to unidentified users or destinations, a security accident can occur.

Figure 1 shows a situation in which secure data is accessed without permission or leaked to an unidentified destination. If the information system can block such cases and give warnings against unauthorized data access or information leakage, the system will operate more safely.

To prepare for these security issues about important data, it is required to control accesses to secure data in programs. Data access control refers to a method for controlling instructions accessible to data objects in computer systems by means of security rules for secure data. Access control is applied in various areas such as cybersecurity enhancement [1], access security model improvement measures [2, 5], and IoT security [3, 4]. In this paper, we propose a method for controlling accesses to secure data in programs to protect programs against unauthorized access to data. The proposed method ensures the data safety of programs by providing a mechanism to ensure safe access to secure data in program execution.

Fig. 1 Data access control system for safe software execution



```

DataSecureProgram ::= Program { statement }
statement ::= { statement }
              | variable = exp
              | statement; statement
              | if ( exp ) then statement
                else statement
              | while ( exp ) do statement
              | output (variable, dst)
              | setSecurityLevel(variable, SecurityLevel)
exp ::= constant | true | false
      | variable
      | exp b exp // binary operation
      | u exp // unary operation
      | input (data, src)
SecurityLevel ::= Secure | Normal | Unknown

```

Fig. 2 The design of data secure language

3 An Approach to Controlling Access

In this section, we design a data secure language and present an approach to controlling access to secure data in a program.

3.1 Data Secure Language

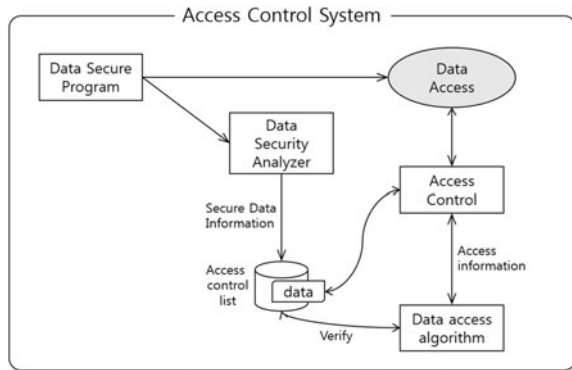
To control accesses to data in a program, we first design a data secure language based on the syntax of procedural programming language WHILE [6], which includes the syntax of imperative programmings, such as sequences, iterations, and conditional branches. Figure 2 shows the syntax of the data secure language designed in this paper. The entire data secure program consists of one or more statements, and the statements include assignments, if statements, while statements, and expressions used for computations in programs.

The setSecurityLevel is a command for specifying the security level of data to control the accesses to secure data. This command sets the security levels of data according to the security importance of the data, such as Secure and Normal. With this command, we can also change the security level of data.

3.2 The Structure of Access Control System for Secure Language

In this section, we design the structure of the access control system for the data secure language. The access control system consists of an access control list [7] and

Fig. 3 The structure of the access control system for data secure programs



an access control module that verifies whether data access is allowed in programs. The access control system permits direct accesses to secure data only if the access is allowed from the data access algorithm.

The access control list is a data structure that manages the security information of data used in programs. The access control system analyzes the statements of programs to verify and control accesses to secure data through the access control rule with the access control list. The data access algorithm determines whether an operation has access right to data and returns the result to the access control module. So, if an operation is determined to have illegal access to data, the operation is blocked and warns users about the illegal access to the data.

Figure 3 shows the structure of the proposed access control system for the data secure language. To ensure secure access to the private information used in programs, the security levels of data are analyzed and managed with the access control list. The access control module verifies access right through the data access algorithm and the access control list according to the access rules for data. In the procedure, if data access instruction violates the access rules to secure data, the system blocks the access and displays a warning message to users.

4 Experimental Evaluation

In this section, the proposed access control system for the data secure language is implemented to evaluate the correctness of the method. The access control system is implemented with the functional language Haskell [8] on the Microsoft Windows 7 operating system. The Haskell has an effective programming environment for designing and developing analyzers for programming languages because it supports various high-level library functions for handling expressions and strings.

With the implementation of the data secure language and access control system, we performed experiments with test programs and evaluate the accuracy of the proposed method. Figure 4 shows the test program for generating secure data and accessing

```

1 Program {
2   dataN = 25;
3   dataS = input(20101225, Normal);
4   setSecurityLevel(dataS, Secure);
5   dataX = input(123999, Unknown);
6   i = 0;
7   num = 2;
8   while (i <= num) {
9     dataS = dataS + i;
10    dataN = i + dataS;
11    dataS = dataS + dataX;
12    i = i + 1;
13  }
14  output(dataN, dataX);
15 }
```

Fig. 4 A benchmark program for evaluating the proposed access control system

the secure data to evaluate the accuracy of the proposed method. The test program has three data, those are dataN for normal data, dataS for secure data, and dataX for dangerous data. In line 10, information of secure dataS is moved to dataN. So, the normal dataN is promoted to having secure information. In line 11, secure dataS and dangerous dataX are computed, and the result is assigned to secure dataS. So, the secure information is contaminated with dangerous information, and this operation is evaluated to be dangerous. In line 14, dataN is moved as output to dangerous dataX. The dataN is initialized as normal data, but it has secure information from the instruction of line 10, so the output of the information to dataX is a dangerous operation.

Figure 5 shows the summarized results of applying the proposed method for the test program. This result displays three warnings for the illegal assignment in line 11, which is executed three times in a while loop. The last output displays a warning

```

[Warning] Secure data is accessed by Dangerous Data.
Danger Type: [Illegal Data Access]
ID: [*] Value: [20101225] Security: Secure
ID: [*] Value: [123999] Security: Unknown
[Warning] Secure data is accessed by Dangerous Data.
Danger Type: [Illegal Data Access]
ID: [*] Value: [20101226] Security: Secure
ID: [*] Value: [20225224] Security: Danger
[Warning] Secure data is accessed by Dangerous Data.
Danger Type: [Illegal Data Access]
ID: [*] Value: [20101228] Security: Secure
ID: [*] Value: [40326450] Security: Danger
[Warning] Secure data outputs to Dangerous Data.
Danger Type: [Illegal Data Output]
ID: [dataX] Value: [60427678] Security: Danger
ID: [dataA] Value: [20101230] Security: Secure
```

Fig. 5 The summarized evaluation results of the access control system for the benchmark program

for the dangerous output operation in line 14. From the evaluation results, the illegal accesses to secure data are correctly identified and blocked by the proposed method. In addition, the proposed method can successfully detect access to secure data that is promoted as secure data during the execution of the program although it is initialized as normal data.

In this paper, we presented a method for controlling access to data in a software system. This proposed method can keep and protect secure data safely in programs through the access control system. The proposed method was implemented and evaluated to verify whether the method can detect and prevent illegal access to secure data effectively.

5 Conclusion

In a recent information system, a vast amount of information is produced and applied in applications. So, the safe management of information is essential for a secure information system. If important information is leaked to a dangerous destination, serious damage may occur. In this paper, we propose an approach to improving software security by controlling accesses to secure data from unauthorized access in programs. We design a data secure language and access control system that can analyze and identify accesses to secure data without authorization. From the evaluation results, we show that the proposed method can effectively detect accesses to secure data in a secure program.

Recently, data usage is growing exponentially, and the technology for managing data is essential for a safe information system. The proposed method can be applied in various areas to maintain data securely in a software system. It is also expected that the proposed approach can be applied in the design and implementation of secure systems that require secure data management in software.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Education) (No. NRF-2017R1D1A1B03034769).

References

1. Thion R (2007) Access control models. Chapter 37. Cyber warfare and cyber terrorism
2. Zhang CN, Yang C (2002) Information flow analysis on role-based access control model. *Inf Manag Comput Secur* 10(5):225–236
3. Cruz-Piris L, Rivera D, Marsa-Maestre I, de la Hoz E, Velasco JR (2018) Access control mechanism for IoT environments based on modelling communication procedures as resources. *Sensors* 18(3)
4. Adda M, Abdelaziz J, Mcheick H, Saad R (2015) Toward an access control model for IOTCollab. *Procedia Comput Sci* 52:428–435

5. Marinovic S, Craven R, Ma J, Dulay N (2011) Rumpole: a flexible break-glass access control model. In: The 16th ACM symposium on access control models and technologies, pp 73–82
6. Bansal AK (2013) Introduction to programming languages. Chapman and Hall
7. Benantar M (2006) Access control systems: security, identity management and trust models. Springer
8. Programming language Haskell. <https://www.haskell.org/>

Implementation of a Container-Based Interactive Environment for Big-Data Analysis on Supercomputer



Seungmin Lee, Ju-Won Park, Kimoon Jeong, and Jaegyoong Hahm

Abstract In this work, we present an environment able to support users who perform big data analysis using distributed and parallel framework to web applications. JupyterHub and Jupyter Enterprise Gateway were used to develop user code in web environment, and Apache Spark is applied as a distributed and parallel framework. The spark cluster deployed at runtime works with Kubernetes as resource management application to maximize the use of resources on the backend and hence all components are container-based. We install all these customized components one of the largest supercomputer, fifth generation supercomputer, NURION, of KISTI. LDAP authenticator plugin and hostPath type volumes are employed to authenticate users of supercomputer and to bind storage respectively. This allows users to perform spark-based big data analysis on the supercomputer through the web interface with interactive environment.

Keywords Big data analysis · Spark · Jupyter · Supercomputer

1 Introduction

Big data analysis is one of the biggest issue together with artificial intelligence, machine learning and deep learning. To support big data analysis, large-scale and scalable infrastructure is required such as a supercomputer and also data analysis is performed by human in the loop processes. The supercomputer environment,

S. Lee (✉) · J.-W. Park · K. Jeong · J. Hahm
Korea Institute of Science and Technology Information, Daejeon 34141, Republic of Korea
e-mail: smlee76@kisti.re.kr

J.-W. Park
e-mail: juwon.park@kisti.re.kr

K. Jeong
e-mail: kmjeong@kisti.re.kr

J. Hahm
e-mail: jaehahm@kisti.re.kr

however, tailored to the traditional scientific workloads are not suitable for big data analysis that shows this usage pattern. To meet this requirement, supercomputing centers provide interactive supercomputing environment [1–3].

In this work, we present an environment able to support users who perform big data analysis using distributed and parallel framework, Apache Spark, to web applications, especially the environment is deployed as a container-based approach. To our knowledge the container-based service for big data analysis that using spark framework on supercomputer has not been studied.

2 Methodology

JupyterHub [4] and Enterprise gateway [5] were used to develop user code in web environment, and Apache Spark [6] is applied as a distributed and parallel framework. The spark cluster deployed at runtime works with Kubernetes [7] as resource management application to maximize the use of resources on the backend and hence all components are container-based. We install all these customized components one of the largest supercomputer, fifth generation supercomputer, NURION, of KISTI.

Figure 1 shows a high-level overview of the JupyterHub and integration at NURION system. The Proxy, Hub, and Pods should be properly configured to integrate user accounts for adherence to the principle of least privilege to mitigate problems caused by the characteristics of docker container. And also user should be able to access data in NURION storage that consists of Lustre file system. LDAP authenticator plugin for JupyterHub and hostPath type volume mount are employed to meet these requirements.

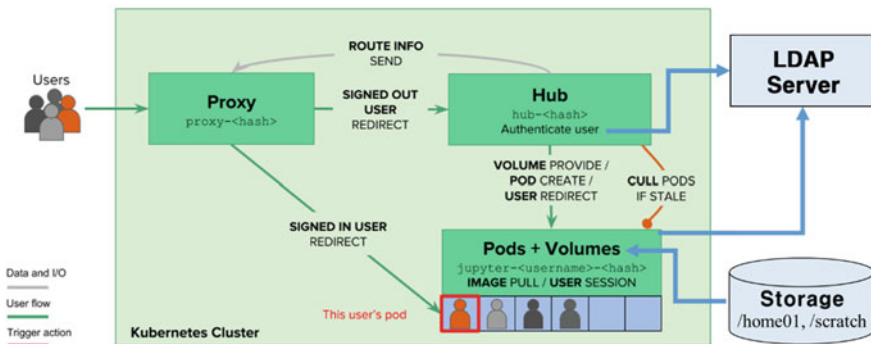


Fig. 1 A high level overview of the JupyterHub and integration at NURION system

```

hub:
  extraConfig:
    myconfig: |
      class LDAPAuthInfo(LDAPAuthenticator):
        def get_user_attributes(self, conn, userdn):
          attrs = {}
          if self.usr_info_attrs:
            found = conn.search(userdn, '(objectClass=*)',
                                attributes=self.usr_info_attrs)
            if found:
              attrs = conn.entries[0].entry_attrs_as_dict
      class LDAPAuthInfoUID(LDAPAuthInfo):
        def pre_spawn_start(self, user, spawner):
          auth_state = yield user.get_auth_state()
          if not auth_state:
            return
          spawner.environment['NB_UID']=
              str(auth_state['uidNumber'][0])
      c.JupyterHub.authenticator_class = LDAPAuthInfoUID
      c.LDAPAuthInfo.server_address = ldap_server_ip
      c.LDAPAuthInfo.usr_info_attrs = ['uid', 'uidNumber']

```

The YAML configuration snippet shown in the above code describes the way to retrieve user information such as uid, username from LDAP server and propagate them through Jupyter kernels by adding environment variables. This make it possible for a user to access data in storage with one's own privileges.

3 Results and Discussion

The construction time of spark cluster at runtime is measured to compare the overhead. Four compute nodes in NURION system are used that consist of Intel Xeon Phi 7250, 16 GB high bandwidth memory (HBM), and 96 GB DDR4 memory for each node [8]. And test system for Intel Xeon is set up with Intel Xeon E5-2680 V2 processors, 2 sockets, 10 cores per socket, and for Intel Skylake is set up with Intel Xeon Silver 4114, 10 cores. One spark driver and 4 workers with 16 threads are used to compare the construction time on the same condition.

Figure 2 shows the construction time measured for big data analysis service from user login to the end of spark cluster ready. Total times are split into the 5 components as follows: the time from user login to start notebook server (SRV_START), generating page to display jupyter notebook (LOAD_PAGE), the time from user request to python kernel launch (KERNEL_READY), spark driver pod is ready (SPARK_DRV) and spark workers are ready (SPARK_EXEC). We exclude the time between page loading and user action because it depends on other factors such as network, user behavior. The construction time for spark cluster is worst in KNL system mainly due

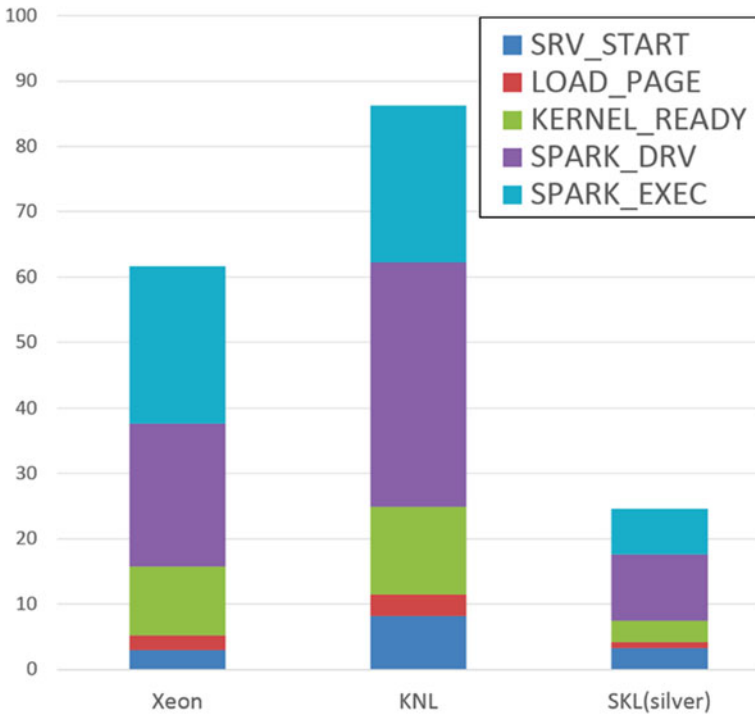


Fig. 2 Time of creating Jupyter server and spark cluster measured in various computing platforms

to lower clock speed of single core, however, KNL has many cores that suitable for distributed and parallel workload. Performance improvement is left future work.

Figure 3 shows the use case scenario of a container-based web interface for big data analysis using spark cluster. User can connect web page to login and use service



Fig. 3 Use case scenario of a container-based web interface for big data analysis using spark

shown in Fig. 3a. Proxy and hub handles user request and create Jupyter notebook server for each user shown in Fig. 3b. And also user can access data in the storage on NURION system or can upload files from local system to the storage. Finally, spark cluster is generated when user selects spark kernel, i.e. Spark—Python (Kubernetes Mode). In this case, there are two workers to perform a simple big data analysis and sample code comes from transactions of a bakery in the Kaggle competition [9]. Note that all processes are in the form of container deployed by Kubernetes that makes system more reliable and flexible.

4 Conclusion

This work explores the container-based interactive environment for big data analysis on supercomputer. JupyterHub, Jupyter Enterprise Gateway, and Apache Spark were employed and customized to integrate with supercomputer. The overhead of constructing spark cluster at runtime exists, however, user can use one's own spark cluster anytime on the web environment.

References

1. Thomas R, Canon S, Cholia S, Gerhardt L, Racah E (2017) Toward interactive supercomputing at NERSC with Jupyter. In: Cray User Group (CUG) conference proceedings, CUG
2. Gobbert JH, Kreuzer T, Grosch A, Lintermann A, Riedel M (2018) Enabling interactive supercomputing at JSC lessons learned. In: ISC high performance
3. Interactive Supercomputing with Jupyter Lab, Swiss National Supercomputing Center [Online]. <https://user.cscs.ch/tools/interactive>
4. JupyterHub [Online]. <https://jupyter.org/hub>
5. Jupyter Enterprise Gateway [Online]. https://jupyter.org/enterprise_gateway
6. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A et al (2016) Apache Spark: a unified engine for big data processing. Commun ACM 56–65
7. Kubernetes [Online]. <https://kubernetes.io>
8. Top500 list [Online]. <https://www.top500.org/list/2019/06>
9. Transactions from a bakery in Kaggle [Online]. <https://www.kaggle.com/sulmansarwar/transactions-from-a-bakery>

Inflight Tracking Method with Beacon System and Scouting Drone



Yunseok Chang

Abstract In this study, we designed an inflight tracking method that can track the current position and the flight route of the target drone through the beacon system and GPS from the scout drone in the beacon range of the target drone. The scouting drone can fly and approach to the inflight target drone to receive the beacon signal in range. If the scout drone receives the beacon signal from a target drone, it checks the beacon information and transmits it to the control center with its GPS data to identify and track the target drone. The inflight tracking method can calculate the current position and flight route of the target drone through the two or more GPS data and flight data at the check positions from the scout drone. The precision of the inflight tracking method can be affected by the critical distance and the number of check positions. To enhance the tracking precision, the scout drone checks two or three different check positions within the critical distance from the inflight target drone. The experimental results showed the proposed inflight tracking method has at least over 92% in the worst case of the experimental environment.

Keywords Inflight tracking · Beacon · Scouting drone · Critical distance · Check position

1 Introduction

Recently, a drone is a kind of very popular device in the area of sports, entertainment, science, and military. Since the drone should be carefully controlled in the air to prevent some kind of accident by a fault or other unpredictable situation, government and regional states have a sort of drone authorization rule. Some kinds of surveillance systems such as the military radar system would be used to detect unauthorized drones [1]. But the radar systems focuses on the aircraft detection that would not be appropriated to the small flight object such as the drone. In the drone area, the

Y. Chang (✉)

Department of Computer Engineering, Daejin University, 1007, Hokook Street, Pocheon, Kyunggi 11159, Republic of Korea

e-mail: cosmos@daejin.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_59

beacon-based drone surveillance system could be more efficient with less resource and cost [2]. But the existing drone surveillance system has short-range compare to the commercial radar system target to the aircraft and limited to the narrow area control at all [3]. Since the beacon system needs power source proportional to the range and a drone has to carry the built-in power source, the effective range of the beacon system also limited to the drone payload [4].

If we want to apply a kind of long-range beacon system, we have to carry a long-range beacon to reach the control center from the air in a wide area. Most of the commercial low power beacon systems can only cover the range of less than 150 m that could not become an effective control system without a high-expensive radar system. In this work, we proposed a kind of bridge or repeater drone called a scout drone between the control center and target drone. The scouting drone relays the beacon information of the target drone that can expand the drone surveillance range. Since the scout drone is also a commercial drone, we can easily implement a kind of bridge between the target drone and the control center as long as we can. If we use a very long-range communication system between the relay drone and the control center, we just approach the relay drone near the target drone to get the beacon information.

2 Inflight Tracking Method

The inflight drone tracking method is a very simple method to identify the remote target drone by using a scout drone that comes from the ground control center. To identify an inflight drone, the target drone has to carry an authorized beacon registered to the control center. By using the scout drone, the control center can identify the flight authority of the target drone from the out of a beacon range, and even can track the inflight target drone through the GPS data from the tracking scout drone by using the very simple method as known as triangulation [5].

Figure 1 shows a scout drone approaching to the target drone. The scouting drone would fly around the target to check its aerial position and the azimuth to the target at two or more positions. From the aerial position data and angles, we can easily calculate the distance between the scout drone and the target drone that can identify the aerial position of the target drone. We use the GPS of the scout drone as the aerial position data for triangulation. Since most of the drone system program libraries provide the GPS API and triangulation method in their development kit, we can easily estimate the distance from the relay drone on the specific position, and identify the aerial position of the target drone without any software modification of the target drone [6, 7].

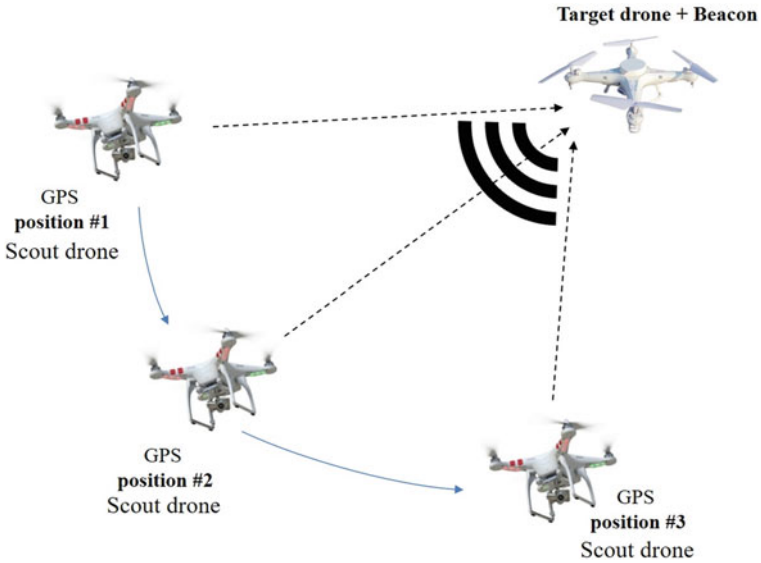


Fig. 1 The basic concept of the inflight tracking for a target drone with beacon

3 Inflight Tracking System Design

Figure 2 shows the target drone that has a commercial long-range beacon and the tracking app. When the tracking app catches the beacon signal from the target drone, we can recognize the target drone as an authorized drone that sends the beacon information of the target drone and current GPS data of the scout drone.

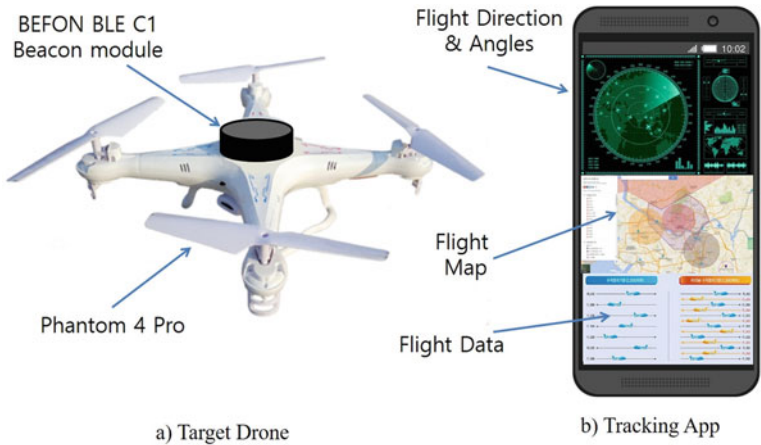


Fig. 2 The target drone with beacon and the tracking app

Table 1 Inflight tracking system environments

System	Parameters	Specifications
Scout drone	Product	DJI phantom 4 advanced
	Flight speed	S/A/P-mode: 72/58/50 km/s
	Weight	1,394 g (with receiver)
Target drone	Product	DJI phantom 4 pro
	Flight speed	S/A/P-mode: 65/58/50 km/s
	Weight	1,392 g (with beacon)
Beacon	Product	BeaFon BLE C1
	Range (Max)	150 m
	ID processing	10 ms

In this work, we made a series of experiments and check two major experimental results, *Critical Distance* and *Tracking Precision* to examine the usability and the efficiency of the proposed methods. Therefore, we tried several experiments with a scout drone and a target drone that has a beacon attached to the drone. Table 1 shows all system environments in the experimental trials.

3.1 Critical Distance

The critical distance $D_{critical}$ can be defined as an average distance from the scout drone to the target drone. If the scout drone approaches the target drone, the scout drone can catch the beacon signal. Since the scout drone always flies to the target drone straight, we can get the GPS data of the scout drone at least two check positions with a different angle, we can estimate the $D_{critical}$ from the GPS data, flight angle and speed of the scout drone by the triangulation directly.

3.2 Tracking Precision

The tracking precision can be defined as a match percentile between the real aerial position and the estimated aerial position calculated by GPS data and direction of the scout drone within the $D_{critical}$. The estimated aerial position of the target drone is calculated at the ground control center. If the estimated aerial position is closer to the real aerial position of the target drone, the tracking method has high precision and reliability.

4 Experimental Results

At first, we checked the critical distance $D_{critical}$ along with the number of check positions and flight speeds of the target drone. Table 2 shows the critical distances at S-Mode and P-Mode flight speed (about 78 and 50 km/h) for a target drone with 2 check positions and 3 check positions at every 8 trials.

Although we had tried to maintain the other experimental condition despite the wind speed and temperature variation, the critical distance has a little deviation for every trial. The experimental results show that the scout drone can catch the beacon signal almost at the same distance independent of the number of the check positions. At the S-Mode and A-Mode flight speed, the critical distance has almost the first-order relation to the flight speed.

We also check the tracking precision for a target drone at 3 flight speed mode with 2 and 3 check positions. Figure 3 shows that 3 check positions have higher precision than 2 check positions. At the P-Mode flight speed, the tracking algorithm shows over 97% tracking precision for the target drone with 3 check positions. Even in the worst case of the tracking method has more than 92% of precision at the S-Mode flight speed. Although the tracking algorithm has less precision at the cases of higher flight speed (S-Mode and A-Mode), these results can show the fact that the scout drone can easily track the inflight drone in the range of its critical distance. If the scout drone cannot recognize any authorized beacon signal from a target drone within the critical distance, the target drone could be recognized as an unauthorized drone. Therefore, the proposed method can be worked as an effective drone tracking system without any expensive airplane or radar system within the scout drone flight range.

Table 2 Critical distances $D_{critical}$ at S-mode/P-mode flight speed with 2/3 check positions

Trials	S-mode	P-mode	S-mode	P-mode
	2 check positions	2 check positions	3 check positions	3 check positions
1	120	126	117	123
2	101	105	104	111
3	92	97	102	101
4	108	111	102	106
5	114	121	110	112
6	101	108	100	103
7	97	101	97	102
8	108	107	102	108

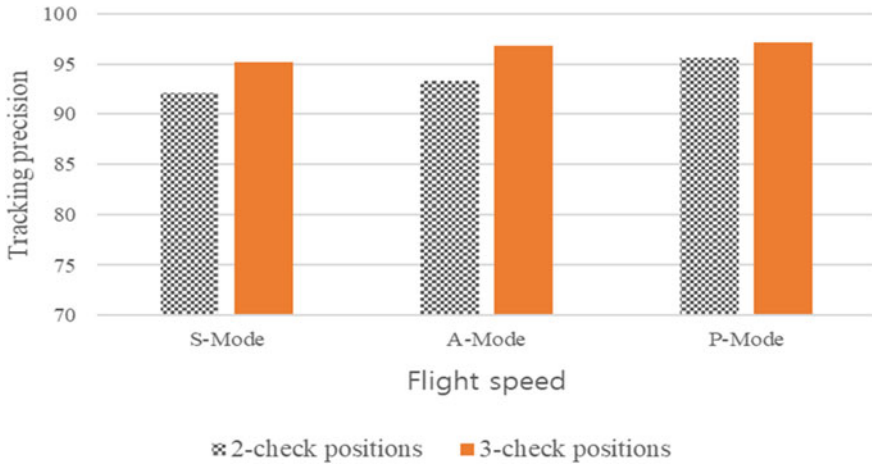


Fig. 3 The tracking precision versus check positions and flight speed

5 Conclusion and Further Work

The inflight tracking method proposed an effective solution to track the flight route and position for a UAV. A scout drone is a sort of long-range beacon bridge solution instead of an active radar or any military system. In this work, we designed an effective inflight drone tracking method with a beacon system on the target drone and the scout drone. The scouting drone checks the beacon signal flying near the target drone. If there is a beacon signal from the target drone within the specific distance, we define it as a critical distance between the target drone and the scout drone. Within the critical distance, we can recognize that the target drone has an authorized beacon or not. Since the critical distance could be varied according to the drone flight speed and positions approaching, we can select the appropriate method to identify the target drone depends on the cost and environment. If we match the critical distance under the specific environment, the experimental results show that the proposed method has over 94 and 97% of tracking precision compare to the real GPS of target drone with two and three check positions. These experimental results show the fact that the proposed method can be applied to the inflight drone applications enough to cover the long-range scouting.

Since the proposed method can not only check the beacon information but also check the aerial position by using the two or more GPS data of the scout drone, the tracking method can be applied in the wide-area such as a regional drone traffic control, a drone hunting or jamming to an illegal inflight drone, and so on. Although the experimental results showed a little bit compromised, we are going to enhance the robustness of the tracking method and beacon system at further work.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2017R1D1A1B03034804).

References

1. Dijkshoom N (2012) Simultaneous localization and mapping with the AR drone. Master's thesis, University of Amsterdam
2. Kendoul F (2012) Survey of advanced in guidance, navigation, and control of unmanned rotorcraft systems. *J Field Robot* 29:315–378
3. Kim B, Jung M, Chang Y (2018) Inflight drone re-routing method with google map on smart pad. In: *Proceedings from ESCS'18*. ACSE, pp 37–40
4. Mac T, Copot C, De Keyser R, Ionescu C (2018) The development of an autonomous navigation system with optimal control of a UAV in a partly indoor environment. *Mechatronics* 49:187–196
5. Neemat S, Inggs M (2012) Design and implementation of a digital real-time target emulator for secondary surveillance radar/identification friend or foe. *IEEE Aerosp Electron Mag* 27(6):17–24
6. DJI Co. (2019) DJI ground pro. <https://www.dji.com/ground-station-pro>
7. DJI Co. (2019) Mobile SDK for phantom 4 series. <https://developer.dji.com/Mobile-SDK>

Deep Learning Based Character Recognition Platform in Complex Situations



BoSeon Kang, Seong-Soo Han, You-Boo Jeon, and Chang-Sung Jeong

Abstract Recently, various character recognition techniques are advanced and used a lot in real situations. In the real complex situation, it is difficult to recognize characters because of Korean, English of various fonts and background noise. In this paper, we shall present character bounding box module, character recognition module and background elimination module to solve problems. We shall show that the character recognition platform is used in complex situations.

Keywords Deep learning · Optical character recognition · Character segmentation · Image processing

1 Introduction

Existing OCR techniques extract edge to predict the shape of a character or use template matching based on standard letters for character recognition [1]. Existing techniques have disadvantages. If the character is a special font or is not in the standard character template, it will not be recognized and need to set the situation conditions perfectly. In addition, most OCR technologies can not be used in actual

B. Kang

Visual Information Processing, Korea University, Seoul, Republic of Korea
e-mail: masati@korea.ac.kr

S.-S. Han

Department of Division of Liberal Studies, Kangwon National University, Samcheok, Republic of Korea
e-mail: sshan1@kangwon.ac.kr

Y.-B. Jeon

Department of Computer Software Engineering, Soonchunhyang University, Asan-si, Republic of Korea
e-mail: jeonyb@sch.ac.kr

C.-S. Jeong (✉)

Department of Electrical Engineering, Korea University, Seoul, Republic of Korea
e-mail: csjeong@korea.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_60

435

situations due to various fonts of problems such as Korean, English, and angles. In order to overcome the disadvantages of the existing OCR techniques, many studies have been conducted to improve the performance of OCR techniques using deep learning [2, 3].

There are two main OCR techniques that use deep learning. The former is the extraction of the area of character, and the latter is the recognition of character in the region of interest. In the case of a technique of extracting a character region, it accurately predicts the area where the letter is likely to be in the input image but does not know which character. In the case of the character recognition technique, the characters can be recognized accurately, but if the input image is large, the letter cannot be recognized accurately. It is even more difficult if the image includes signs, notices, and many products. Therefore, in this paper, we shall present the platform available in real complex situations by combining the technique of extracting a character region and the character recognition technique.

The rest of this paper is organized as follows: Sect. 2 presents related works, and Sect. 3 presents a platform architecture, and Sect. 4 presents the experimental result and processing to recognize characters. Finally, Sect. 5 concludes the paper.

2 Related Works

2.1 Character Segmentation

Existing Segmentation-based Text Detection are extracted character candidates and through filtering. Then proceed with grouping. These methods were time-consuming and required post-processing steps [4]. In addition, the unsatisfactory performance made it difficult to apply in real life. However, the performance of the Regression-based Text Detection technique is improved by developing the SSD model [5]. The SSD model uses anchors of various shapes to predict the shape of irregular texts. Thus, speed and accuracy are significantly improved over the existing model.

2.2 Character Recognition

The Scene text recognition model using deep learning is the most important technology, it performs well regardless of some background noise [6, 7]. Before the development of the deep learning model, it was used only in a set environment, such as clean documents. The recognition process of good performance models is in the order of normalizing, feature extraction, sequence modeling, and prediction [8]. In the image normalize the process, it is a process of horizontally spreading diagonal and arched characters and extract features from normalized images except fonts,

sizes, etc. And we predict the characters from the extracted features using Attention Network [9].

2.3 Background Elimination

We use connected components and threshold methods to eliminate the background. We analyze the background of the image and use the connected components method in case of a complex background, and the threshold method in case of a simple background to improve the accuracy of character recognition. The connected components technique uses BBDT [10] and SAUF [11] algorithms to find the connected components for the character in the image and extract only the character area accurately.

3 Platform Architecture

In this section, we present a platform for correct character recognition in complex situations. Its platform architecture is shown in Fig. 1.

The platform consists of Character Bounding Box Module, Character Recognition Module, Background Elimination Module, and Training Module. The Character Bounding Box Module consists of Text/non-text Prediction Function and Link Prediction Function and Prediction Data Analyzer, and Bounding Box Optimizer. The Text/non-text Prediction Function measures the possible values of the text area in the input image. The

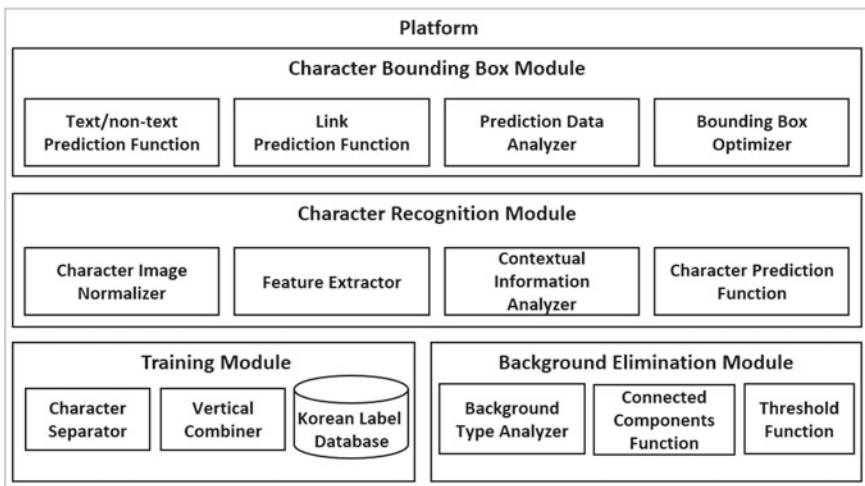


Fig. 1 Platform architecture

value is expressed in pixels and the area is divided by connecting adjacent pixels in the Link Prediction Function. Using the pixel and link method tends to loosen the bounding box. The Bounding Box Optimizer makes the segmented region tight using parameter values.

The Character Recognition Module consists of Character Image Normalizer, Feature Extractor, Contextual Information Analyzer, and Character Prediction Function. Character Image Normalizer receives the processed image from the Bounding Box Optimizer. If the text included in the image is not horizontal, it not be recognized. So we normalize the input image using the spatial transformation network. Feature Extractor extracts features from normalized images. In this step, the shape of the character is inferred through the CNN model. The Contextual Information Analyzer analyzes the extracted features in the previous step. It sends the features to the next level, except those that are not contextually relevant. Finally, Character Prediction Function recognizes characters. It uses the attention network to capture features to predict text.

The Background Elimination Module analyzes the background type for the inputted bounding box area. The background is eliminated by using the connected components function and threshold hold function based on background type.

Finally, the Training Module receives the results of the character bounding box module and split it. The divided characters are recombined into vertical and horizontal words for deep learning training and stored in the Korean label database.

4 Experiment

The sample dataset at the 2019 Artificial Intelligence R&D Grand Challenge is used to evaluate our platform. The challenge sample dataset the FHD (1920 × 1080) resolution. It comprises complex, varied streets, and roads. And the score of the character bounding box is measured using the IOU (Intersection over Union) in this challenge. It is acknowledged as the correct answer if the region of the bounding box of the answer and region of the predicted bounding box are similar over 75%. The character recognition part acknowledges as the correct answer only the exact same character.

4.1 Experiment Process

We conduct experiments in the procedure as Fig. 2. The score is measured at each procedure. Our experiment environment uses CPU Intel i7-7700 3.60 GHz, NVIDIA GeForce GTX 1080, and 32 GB RAM.

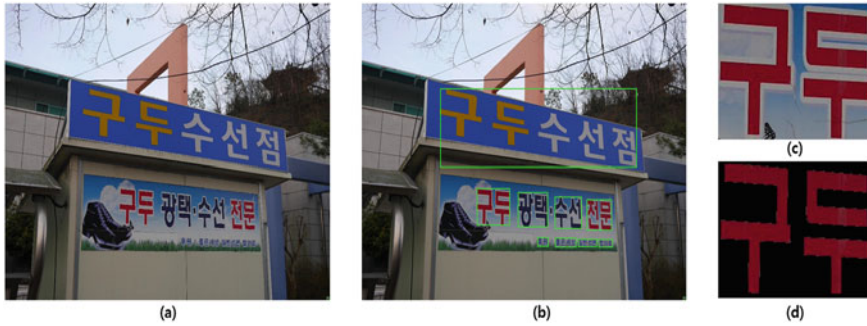


Fig. 2 a Original image, b character segmented image, c character bounding box image, d background eliminated bounding box image

4.2 Experiment Result

The character bounding box score is 68.591 and the character recognition module score is 83.374. The score of the character recognition module proceeding with the background elimination is 84.7. The final score is 63.88. The answer to challenge does not contain special characters. However, special characters are closely related to the characters, so it is difficult to remove them. The factors with low scores are the special characteristic of Korean. Because the number of all cases of Korean characters is 11,172, the whole can not be trained. Deep learning models cannot recognize unlearned Korean characters.

5 Conclusion

In this paper, we have presented a character recognition platform that can be used in real complex situations. The platform used the character bounding box module to extract the character region and recognize the character in the extracted region. We also improved the accuracy by removing the background. We have shown that even in real complex situations, we can use character recognition using deep learning. Experiment results show that improving the performance of the character bounding box can also improve the final score.

Our future work is to study how to unite Korean written from special angles. Even in more complex situations, we should be able to find and recognize the letter area. We will study how to train countless Korean characters efficiently.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03035461), and supported by 2019 Research Grant from Kangwon National University, and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-2014-1-00720),

(IITP-2019-2014-1-00720) supervised by the IITP (Institute for Information & communications Technology Promotion), and the Soonchunhyang University Research Fund.

References

1. Arica N, Yarman-Vural FT (2016) Optical character recognition for cursive handwriting. *IEEE Trans Pattern Anal Mach Intell* 24(6):801–813. [IEEE](#)
2. Yao C, Bai X, Sang N et al (2016) Scene text detection via holistic, multi-channel prediction. [arXiv:1606.09002](#)
3. He T, Huang W, Qiao Y, Yao J (2016) Accurate text localization in natural image with cascaded convolutional text network. [arXiv:1603.09423](#)
4. Deng D, Liu H, Li X, Cai D (2018) Pixellink: detecting scene text via instance segmentation. [arXiv:1801.01315](#)
5. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2015) SSD: single shot multibox detector. [arXiv:1512.02325](#)
6. Shi B, Bai X, Yao C (2016) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans Pattern Anal Mach Intell* 39:2298–2304
7. Shi B, Wang X, Lyu P, Yao C, Bai X (2016) Robust scene text recognition with automatic rectification. In: *Proceedings of CVPR*, pp 4168–4176
8. Baek J (2019) What is wrong with scene text recognition model comparisons? Dataset and model analysis. [arXiv:1904.01906](#)
9. Cheng Z, Bai F, Xu Y, Zheng G, Pu S, Zhou S (2017) Focusing attention: towards accurate text recognition in natural images. In: *Proceedings of ICCV*, pp 5076–5084
10. Grana C, Borghesani D, Cucchiara R (2010) Optimized block-based connected components labeling with decision trees. *IEEE Trans Image Process* 19(6):1596–1609
11. Bolelli F, Cancilla M, Grana C (2017) Two more strategies to speed up connected components labeling algorithms. In: *Battiatto S, Gallo G, Schettini R, Stanco F (eds) Proceedings of ICIAP 2017*. LNCS, vol 10485. Springer, Cham, pp 48–58

Design of Restricted Coulomb Energy Neural Network Processor for Multi-modal Sensor Fusion



Jaechan Cho, Minwoo Kim, Yongchul Jung, and Yunho Jung

Abstract This paper proposes a restricted coulomb energy neural network (RCE-NN) with an improved learning algorithm and presents the hardware architecture design and FPGA implementation results. The proposed learning algorithm divides each neuron region in the learning process and measures the reliability with different factors for each region. In addition, it applies a process of gradual radius reduction by a pre-defined reduction rate. In performance evaluations using two datasets, RCE-NN with the proposed learning algorithm showed high recognition accuracy with fewer neurons compared to existing RCE-NNs. The proposed RCE-NN processor was implemented in an Intel-Altera Cyclone IV FPGA with 26,702 logic elements, 13,096 registers and 131,072bits memory and operated at the clock frequency of 150 MHz.

Keywords Artificial neural network (ANN) · FPGA · Machine learning · Pattern recognition · Restricted coulomb energy neural network (RCE-NN)

1 Introduction

Intelligent multi-modal sensor fusion is the most popular technology in various fields such as human machine interfaces (HMI), speech recognition, robotics and medical applications [1]. Moreover, an artificial neural network (ANN) algorithm has become

J. Cho · M. Kim · Y. Jung · Y. Jung (✉)

School of Electronics and Information Engineering, Korea Aerospace University, Goyang-si, South Korea

e-mail: ycjung@kau.kr

J. Cho

e-mail: jccho@kau.kr

M. Kim

e-mail: minwoo@kau.kr

Y. Jung

e-mail: yjung@kau.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_61

a key technology to optimize data-driven approaches of multi-modal sensor fusion. Deep neural networks (DNNs) are a part of the broad field of ANN and deliver state-of-the-art accuracy on many ANN tasks [2]. However, to complete the tasks with higher accuracy, DNNs take huge computing resources and can take several days depending on the size of the dataset and the number of layers in the network. In addition, since most DNNs are designed for specific application fields, they have a fixed network structure such as the fixed number of layers and neurons. Therefore, it is not suitable for multi-modal sensor fusion applications requiring different network structures depending on the characteristics of feature data of each sensors in the learning process.

In contrast, the restricted coulomb energy neural network (RCE-NN) can flexibly modify the network structure because it generates new neuron only when necessary [3–5]. Therefore, it can support various multi-modal sensor fusion applications and has recently been implemented for various embedded systems used in HMI. The RCE-NN efficiently classifies feature distributions by constructing hyperspherical neurons with radii and hypersphere centers. However, learning schemes of existing RCE-NN applies an inefficient radius adjustment, such as learning all neurons at the same radius or reducing the radius excessively in the learning process. Moreover, since the reliability of eliminating unnecessary neurons is estimated without considering the activation region of each neuron, it is inaccurate and leaves unnecessary neurons extant.

In this paper, an efficient learning algorithm for RCE-NN is proposed. The reliability of each neuron is estimated by considering the activation region with different factors. In addition, the radius is gradually reduced at a pre-defined reduction rate to prevent the generation of unnecessary neurons. The design and implementation results of the RCE-NN processor for real-time processing are also presented.

2 Restricted Coulomb Energy Neural Network

The RCE-NN consists of an input layer, a prototype layer (hidden layer) and an output layer. The input layer comprises feature vectors and all feature vectors are connected to each neuron of the prototype layer [3]. The prototype layer is the most essential part of the RCE-NN. Neurons in this layer save hypersphere centers and radii, which construct hyperspherical classifiers in the feature space. The output layer uses the neuron's response to output the label value of the neuron that best matches the input feature vector.

Each neuron \mathbf{n}_j in the prototype layer contains the information as follows:

$$\mathbf{n}_j = [c_j^1, c_j^2, \dots, c_j^k, r_j, l_j, z_j], \quad (1)$$

where $j \in \{1, 2, 3, \dots, m\}$ is the neuron index and the total number m varies according to the learning results. That is, if m neurons are learned after learning is completed,

they are defined as a set of neurons $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_m]$. Each neuron \mathbf{n}_j contains a hypersphere center $\mathbf{c}_j = [c^1_j, c^2_j, \dots, c^k_j]$, radius r_j , activation count z_j and learned label l_j , where k is the number of features in the input feature vector used in the learning [5].

If the number of input feature vectors during the learning process is w , the input feature vector set can be represented by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_w]$, where the feature vector \mathbf{x}_i consists of k features and a label l_x . The feature vector \mathbf{x}_i is entered to each neuron and the distance between the feature vector \mathbf{x}_i and the hypersphere center \mathbf{c}_j is computed as follows:

$$d(\mathbf{x}_i, \mathbf{c}_j) = \sqrt{(x^1_i - c^1_j)^2 + (x^2_i - c^2_j)^2 + \dots + (x^k_i - c^k_j)^2}. \tag{2}$$

Then, neuron \mathbf{n}_j is activated only if $d(\mathbf{x}_i, \mathbf{c}_j) \leq r_j$. If no neurons are activated for the feature vector \mathbf{x}_i , a new neuron \mathbf{n}_{n+1} with a hypersphere center $\mathbf{c}_{n+1} = [x^1_i, x^2_i, \dots, x^k_i]$, label l_x , and radius R is generated, where $[x^1_i, x^2_i, \dots, x^k_i]$ and l_x are the feature values and label of the feature vector \mathbf{x}_i , respectively, and R is the pre-defined global radius. In addition, the total number of neurons m is increased by one.

3 Proposed RCE-NN Processor

3.1 Proposed Learning Algorithm

In order to overcome the problems of existing learning schemes, the proposed learning algorithm estimates reliability by dividing the activation region associated with each neuron and increasing the z_j with different factors for each region in the learning process as follows:

$$z_j^{t+1} = \begin{cases} z_j^t + 1, & \left(\frac{r_j + R_{min}}{2}\right) < d(\mathbf{x}_i, \mathbf{c}_j) \leq r_j \\ z_j^t + \alpha, & R_{min} < d(\mathbf{x}_i, \mathbf{c}_j) \leq \left(\frac{r_j + R_{min}}{2}\right), \\ z_j^t + \beta & d(\mathbf{x}_i, \mathbf{c}_j) \leq R_{min} \end{cases}, \tag{3}$$

where α and β are experimentally determined according to the distribution of the feature vector ($1 < \alpha < \beta$). That is, when the current input feature vector activates a neuron in a region near the hypersphere center, the z_j is increased with a higher factor. When it activates a neuron in the boundary area, the z_j is increased with a lower factor. In addition, when a neuron with a different label from that of the input feature vector is activated, the proposed algorithm gradually decreases the radius

Table 1 Performance evaluation results ($\alpha = 5, \beta = 10$)

		TR [3]	HPL [4]	DDA [5]	Proposed
GAS [6]	Number of neurons	509	559	1,026	449
	Recognition accuracy	84.31%	91.29%	95.17%	96.86%
MCHP [7]	Number of neurons	752	911	1,164	641
	Recognition accuracy	89.65%	95.32%	97.38%	98.52%

according to reduction rate γ . This not only improves the recognition accuracy but also suppresses the generation of unnecessary neurons.

We conducted learning and recognition tasks with all methods for RCE-NN on two datasets for a gas sensor and motion-capture hand postures (MCHP) [6, 7]. Learning and recognition were performed by selecting 7,000 learning samples and 6,910 test samples randomly from the whole sample. As a result of the performance evaluation as shown in Table 1, the RCE-NN with the proposed learning algorithm shows good recognition accuracy with half the number of neurons that RCE-NN with a dynamic decay adjustment (DDA) used [4]. In addition, the proposed algorithm shows better recognition accuracy from 3 to 12% with fewer neurons than RCE-NN with traditional (TR) [3] and a hierarchical prototype learning (HPL) [5].

We also constructed a 3D number dataset which was generated by extracting the accelerometer values at a sampling rate of 20 Hz. Five participants were asked to hold the inertial measurement (IMU) sensor to write ten digits in the air. Each participant wrote each digit 20 times, and data corresponding to a total 1000 hand gestures were collected. We employed the hand gestures recorded from five participants, and the evaluation results were obtained by fivefold cross-validation. As a result, the proposed RCE-NN showed the 98.6% average recognition accuracy, which outperformed the others for all users.

3.2 *Hardware Architecture Design and Implementation Results*

Figure 1 shows the block diagram of the proposed RCE-NN processor, including a feature memory unit (FMU), neuron unit (NU), activated neuron detection unit (ANDU), and network control unit (NCU). In the learning process, the unlearned NUs store the feature vectors from FMU in the neuron memory, and the learned NUs calculate distance by (2). Then, the learned NUs compare the distance to the stored radius, and the activated NUs send the distance and label to the ANDU. The ANDU analyzes the output of the activated NUs and transfers the information of the minimum distance and label to the NCU. NCU determines whether to generate a new NU or remove an existing NU.

The proposed RCE-NN processor was designed in Verilog hardware description language (HDL) and implemented in an Intel-Altera Cyclone IV FPGA to verify that

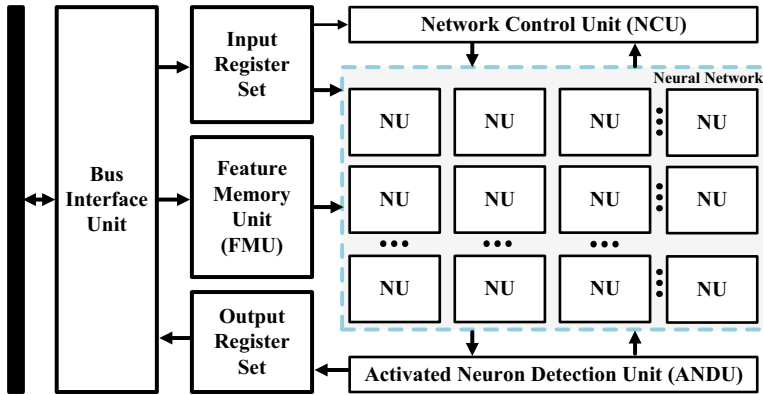


Fig. 1 Block diagram of the proposed RCE-NN processor

Table 2 Key features of the implemented RCE-NN processor

Parameter	Value
Device	Intel-altera cyclone IV
Logic elements	26,702
Registers	13,096
Internal memory	131,072 bits

real-time learning and recognition was possible. It is observed that the proposed architecture requires 26,702 logic elements, 13,096 registers and 131,072bits memory as shown in Table 2. In addition, we confirmed that the real-time learning and recognition were possible because the proposed RCE-NN possible required only 0.93 μ s for learning and 0.96 μ s for recognition, at an operating frequency of 150 MHz.

4 Conclusion

In this paper, we proposed an efficient RCE-NN processor with an improved learning algorithm. Learning algorithms in existing RCE-NNs show degraded recognition performance and increased complexity because of the inaccurate reliability of learned neurons and inefficient radius adjustments. To overcome this problem, the proposed algorithm divides the activation region of each neuron in the learning process and measures the reliability with different factors for each area, and gradually reduces the radius using a pre-defined rate. In performance evaluation using two datasets, RCE-NN with the proposed learning algorithm showed good recognition accuracy with fewer neurons compared with existing RCE-NNs. We also designed the hardware for its real time operation. The designed RCE-NN processor has 26,702 logic elements,

13,096 registers and the memory requirement of 131,072bits, and it can support real-time learning and recognition at an operating frequency of 150 MHz.

Acknowledgements This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00056) and CAD tools were supported by IDEC.

References

1. Garcia F, David M, Arturo DE, Jose MA (2017) Sensor fusion methodology for vehicle detection. *IEEE Trans ITSM* 9(1):123–133
2. Sze V, Chen Y, Yang T, Ember J (2017) Efficient processing of deep neural networks. *Proc IEEE* 105:2295–2329
3. Hudak M (1992) RCE Classifiers: theory and practice. *Cybern Syst* 23:483–515
4. Berthold M, Diamond J (1995) Boosting the performance of RBF networks with dynamic decay adjustment. *Adv Neural Inf Process Syst (NIPS 7)*, 521--528
5. Sui C, Kwok N, Ren T (2011) A restricted coulomb energy (RCE) neural network system for hand image segmentation. In: *IEEE conference on computer and robot vision*. St. Johns, pp 270--277
6. Center for machine learning and intelligent systems. UCI machine learning repository. <https://archive.ics.uci.edu/ml/datasets/gas+sensor+array+drift+dataset>
7. Center for machine learning and intelligent systems. UCI machine learning repository. <https://archive.ics.uci.edu/ml/datasets/Motion+Capture+Hand+Postures>

Low Complexity Pipelined FFT Processor for Radar Applications



Yongchul Jung, Jaechan Cho, and Yunho Jung

Abstract This paper proposes a low complexity fast Fourier transform processor for radar applications based on the radix- 2^2 and the radix- 2^3 single-path delay feedback pipeline architectures. The delay elements for aligning the data in the pipeline stage are one of the most complex units, and that of stage 1 is the biggest. By exploiting the fact that the input data sequence is zero-padded and that the twiddle factor multiplication in stage 1 is trivial, the proposed FFT processor can dramatically reduce the required number of delay elements. Moreover, the 256-point FFT processors were designed using hardware description language (HDL) and implemented on a Xilinx Artix-7 FPGA device. The proposed architecture was implemented with 1782 logic slices, which can be efficient and suitable for zero-padded FFT processors.

Keywords Delay elements · Fast fourier transform (FFT) · Single-path delay feedback (SDF) · Zero-padded signal

1 Introduction

The fast Fourier transform (FFT) is a mathematical algorithm for reducing the computational complexity of the discrete Fourier transform (DFT) and is widely used for frequency analysis. In radar applications, high frequency resolution is required to measure the exact position of the target. The zero-padded FFT offers increased frequency resolution by extending the length of the input data sequence in the time domain by padding with zeros at the tail of the discrete time signal. Because of this, the zero-padding method increases the complexity of the FFT processor and,

Y. Jung · J. Cho · Y. Jung (✉)

School of Electronics and Information Engineering, Korea Aerospace University, Goyang-si, Korea
e-mail: yjung@kau.kr

Y. Jung

e-mail: ycjung@kau.kr

J. Cho

e-mail: jccho@kau.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_62

447

reducing its complexity can notably contribute to a more efficient hardware design [1].

The radix-2 and radix-4 algorithms are the most widely used for implementing FFT processors because of their simple architectures. For pipeline architectures, the radix-4 algorithm has a smaller number of non-trivial multiplications than the radix-2 algorithm. However, the radix-4 algorithm complicates the control of butterfly architectures more than the radix-2 algorithm. Thus, the radix-2² and radix-2³ algorithms have been proposed to reduce the complexity of high-radix algorithms. The radix-2² algorithm has the same number of non-trivial multiplications as the radix-4 algorithm but maintains the butterfly architecture of the radix-2 algorithm. Similarly, the radix-2³ algorithm has the same number of non-trivial multiplications as the radix-8 algorithm [2].

Single-path delay feedback (SDF) pipeline FFT architectures are commonly used because they have the smallest number of non-trivial multiplications compared with other pipeline architectures, such as single-path delay commutator (SDC) and multi-path delay commutator (MDC). However, as the number of FFT points increases, the SDF architecture requires significantly more circuit area because of the delay elements for data reordering [3].

In this paper, we propose an area-efficient FFT processor for zero-padded signals by taking advantage of the fact that the data sequence is zero-padded and that the twiddle factor (TF) operation in stage 1 is a trivial multiplication in the radix-2² and radix-2³ algorithms. The rest of this paper is organized as follows. In Sect. 2, we review the zero-padded FFT. The hardware architecture of the proposed FFT processor is described in Sect. 3. In Sect. 4, we compare the proposed zero-padded FFT architecture with conventional architectures. Finally, Sect. 5 concludes the paper.

2 Zero-Padded FFT

The DFT for complex data sequence $x(n)$ of length N is defined as follow

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk}, \quad 0 \leq n, k \leq N - 1. \quad (1)$$

When analyzing the resolution of the DFT, there are two factors to consider. The first one is the spectral resolution, which refers to the algorithm's capability to detect closely spaced spectral components. The second one is the frequency resolution, which is the definition of the distance between frequency bins. Whereas the spectral resolution can only be increased by increasing the time window of the signal, the frequency resolution is determined by the number of input data points in the sequence given to the DFT. A longer data sequence is usually obtained by using the zero-padding method, which is described below.

Assume that a new data sequence $y(n)$ is created by zero-padding the original data sequence $x(n)$ of length N to a length of M .

$$y(n) = \begin{cases} x(n), & 0 \leq n \leq N - 1 \\ 0, & N \leq n \leq M - 1 \end{cases} \quad (2)$$

The M -points of the DFT are calculated as

$$Y(k) = \sum_{n=0}^{M-1} y(n) W_M^{nk}, \quad 0 \leq k \leq M - 1. \quad (3)$$

Based on the divide-and-conquer algorithm, indices n and k can be written as

$$n = \frac{M}{2}n_1 + \frac{M}{4}n_2 + n_3, \quad (4)$$

$$k = k_1 + 2k_2 + 4k_3, \quad (5)$$

where $0 \leq n_1 \leq 1$, $0 \leq k_1 \leq 1$, $0 \leq n_2 \leq 1$, $0 \leq k_2 \leq 1$, $0 \leq n_3 \leq (M/4 - 1)$, and $0 \leq k_3 \leq (M/4 - 1)$ [4]. Replacing Eqs. (4) and (5) in Eq. (3), we obtain

$$\begin{aligned} Y(k_1 + 2k_2 + 4k_3) &= \sum_{n_3=0}^{M/4-1} \sum_{n_2=0}^1 \sum_{n_1=0}^1 y\left(\frac{M}{2}n_1 + \frac{M}{4}n_2 + n_3\right) W_2^{n_1 k_1} \\ &\quad W_4^{n_2(k_1+2k_2)} W_M^{n_3(k_1+2k_2+4k_3)} \\ &= \sum_{n_3=0}^{M/4-1} \sum_{n_2=0}^1 B_{M/2}^{k_1} \left(\frac{M}{4}n_2 + n_3\right) W_4^{n_2(k_1+2k_2)} W_M^{n_3(k_1+2k_2+4k_3)}. \end{aligned} \quad (6)$$

In Eq. (6), the butterfly operation is given by

$$\begin{aligned} B_{M/2}^{k_1} \left(\frac{M}{4}n_2 + n_3\right) &= \sum_{n_1=0}^1 y\left(\frac{M}{2}n_1 + \frac{M}{4}n_2 + n_3\right) W_2^{n_1 k_1} \\ &= y\left(\frac{M}{4}n_2 + n_3\right) + (-1)^{k_1} y\left(\frac{M}{2} + \frac{M}{4}n_2 + n_3\right). \end{aligned} \quad (7)$$

Assuming that M is $2N$ in order to increase the frequency resolution twice, samples from $y(N)$ to $y(2N - 1)$ are set to zero so that Eq. (7) can be simplified as follows:

$$B_{M/2}^{k_1} \left(\frac{M}{4} n_2 + n_3 \right) = y \left(\frac{M}{4} n_2 + n_3 \right). \quad (8)$$

Therefore, Eq. (6) can be summarized as follows

$$\begin{aligned} Y(k_1 + 2k_2 + 4k_3) &= \sum_{n_3=0}^{M/4-1} \sum_{n_2=0}^1 (-j)^{n_2 k_1} y \left(\frac{M}{4} n_2 + n_3 \right) W_2^{n_2 k_2} W_M^{n_3(k_1+2k_2+4k_3)} \\ &= \sum_{n_3=0}^{M/4-1} H(k_1, k_2, n_3) W_M^{n_3(k_1+2k_2)} W_{M/4}^{n_3 k_3} \end{aligned} \quad (9)$$

where the output of the stage-2 butterfly $H(k_1, k_2, k_3)$ is expressed as shown in Eq. (10):

$$\begin{aligned} H(k_1, k_2, n_3) &= \sum_{n_2=0}^1 (-j)^{n_2 k_1} y \left(\frac{M}{4} n_2 + n_3 \right) W_2^{n_2 k_2} \\ &= y(n_3) + (-1)^{k_2} (-j)^{k_1} y \left(\frac{M}{4} + n_3 \right). \end{aligned} \quad (10)$$

3 Proposed Hardware Architecture

In order to double the frequency resolution, the tail of input data sequence $x(n)$ of length N is padded with N zeros to double its length in the time domain. The FFT signal flow graph (SFG) of the radix-2² algorithm for a zero-padded signal with double frequency resolution is shown in Fig. 1. To implement the zero-padded FFT using the conventional radix-2² SDF architecture, delay elements of length N are required for data sequence reordering in stage 1, and the length of the delay elements required for each stage is reduced by half each time as shown in Fig. 2. That is, in order to implement the FFT processor for a zero-padded signal of length $2N$ using the conventional radix-2² SDF architecture, delay elements with a total length of $2N-1$ are required [4]. As a result, the number of delay elements notably increases with the FFT data points. To solve this problem, we propose the hardware architecture depicted in Fig. 3 by using the feedback path of the SDF architecture and exploiting the trivial multiplication of stage 1.

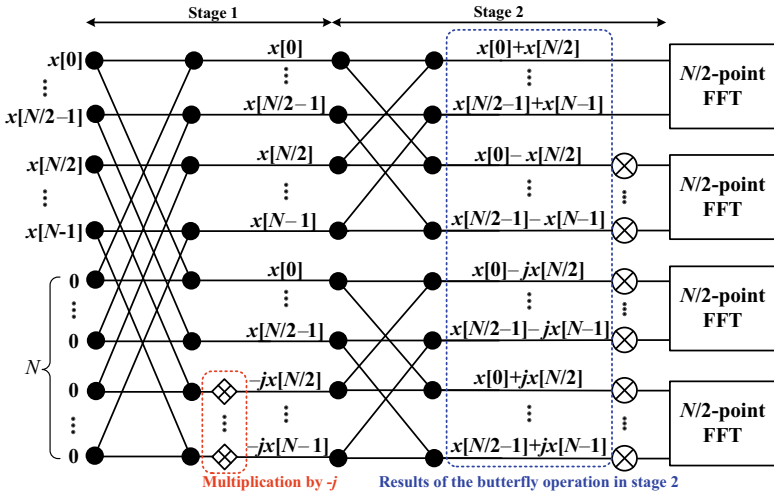


Fig. 1 Signal flow graph for double frequency resolution

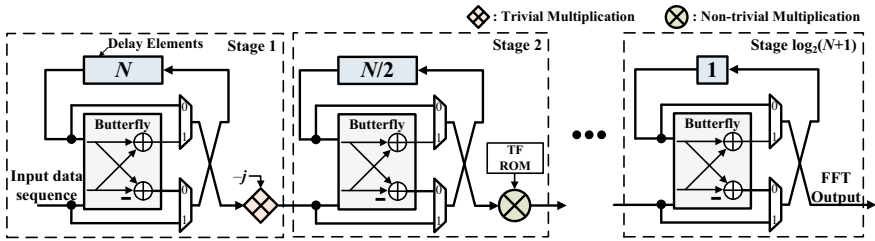


Fig. 2 Hardware architecture of the conventional SDF FFT processor

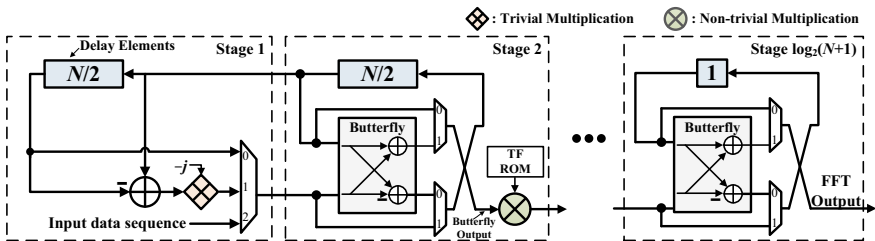


Fig. 3 Hardware architecture of proposed SDF FFT processor for double frequency resolution

4 Comparison

Table 1 shows a comparison of the hardware area and performance between the conventional pipelined FFT architecture and the proposed hardware architecture for

Table 1 Comparison of pipeline hardware architectures for the computation of a 2^q -point zero-padded FFT on complex-valued data

Pipelined architecture	Complex adders	Complex multipliers	Delay elements	Latency (Cycles)
SDF Radix-2	$2q$	$q-1$	2^q-1	2^q
SDF Radix-4	$4q$	$q/2-1$	2^q-1	2^q
SDF Radix- 2^2	$2q$	$q/2-1$	2^q-1	2^q
SDF Split-radix-2	$2q$	$q/2-1$	2^q-1	2^q
SDC Radix-4	$3q/2$	$q/2-1$	$2^{q+1}-1$	2^q
Proposed SDF Radix- 2^2	$2q-3$	$q/2-1$	$3(2^{q-2})-1$	$3(2^{q-2})$

a zero-padded signal of length 2^q when the double frequency resolution. The table includes the area in terms of complex adders, complex multipliers, and the number of delay elements, as well as latency. Additionally, the number of complex multipliers is the same as in the radix- 2^2 SDF architecture, but it can be seen that the number of complex adders is reduced by 2 compared with the radix- 2^2 SDF architecture. Most notably, compared with the conventional hardware architecture (in which the number of delay elements significantly increases with FFT length and the number of data paths), the proposed hardware architecture reduces the number of the delay elements are significantly. Moreover, latency is significantly reduced compared with other single-path pipeline architectures.

In order to confirm the superiority of the proposed architecture, we implemented two 256-point FFT processors with the proposed and conventional radix- 2^2 SDF architectures. For four-times frequency resolution, the tail of a input data sequence of length 128 is padded with 128 zeros. A 12-bit word for real and imaginary data paths was selected to satisfy the requirement for a signal-to-quantization noise-ratio (SQNR) of 40 dB. Two FFT processors were designed using hardware description language (HDL) and implemented on a Xilinx Artix-7 FPGA device. Table 2 shows comparison of implementation results. As depicted in this Table, the proposed architecture can reduce the logic slices by 22.6% compared to the conventional architecture owing to the reduction of 25.1% for delay elements.

Table 2 Comparison of implementation results of a 256-point double frequency resolution zero-padded FFT on complex-valued data

Block name	SDF Radix- 2^2	Proposed	Reduction (%)
Butterfly unit	128	119	7.0
Non-trivial Multiplier	135	135	0
Delay elements	2040	1528	25.1
Total	2303	1782	22.6

5 Conclusion

In this paper, we proposed an area-efficient FFT processor for zero-padded signals based on the radix-2² and radix-2³ SDF pipeline architectures by taking advantage of the fact that the input data sequence is zero-padded that and the twiddle factor multiplication in stage 1 is trivial. The proposed FFT processor can dramatically reduce the required the number of delay elements. For four-times frequency resolution, the tail of a input data sequence of length 128 is padded with 128 zeros, the number of delay elements can be reduced by 25.1%, and we demonstrated that the proposed architecture is efficient and suitable for zero-padded FFT processors.

Acknowledgements This work was supported by Civil-Military Technology Cooperation Program, 16-CM-RB-12, funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) and Defense Acquisition Program Administration (DAPA, Korea), and CAD tools were supported by IDEC.

References

1. Oppenheim AY et al (2009) Discrete-time signal processing. englewood cliffs, vol 3. NJ, USA, Prentice-Hall
2. Ayinala M, Parhi KK (2010) Parallel-Pipelined Radix-2² FFT architecture for real valued signals. In: Proceedings of asilomar conference signals, systems and computers, pp 1274—1278
3. Yin X, Yu F, Ma Z (2016) Resource-efficient pipelined architectures for Radix-2 real-valued FFT with real datapaths. In: IEEE transaction on circuits system-II: express breifs, pp 803--807
4. He S, Torkelson M (1996) A new approach to pipeline FFT processor. In: Proceedings of international conference on parallel processing. Honolulu, pp 766--770

Dynamic Mitigation of Catastrophic Forgetting Using the Sampling Network



Dae Yong Hong, Yan Li, and Byeong-Seok Shin

Abstract The catastrophic forgetting in transfer learning makes a neural network lose the performance on previously learned datasets when dealing with large amounts of data. Predictive Elastic Weight Consolidation (PEWC) reduces the catastrophic forgetting by extracting only images with relatively more incorrect network predictions, but uses static sampling technique. PEWC also includes images in the training data which can be correctly classified by the network, leaving the possibility for further reduction of the training data. In this paper, we additionally apply a sampling network that extracts images dynamically without sorting, so that only images whose predictions are similarly inaccurate in general are used for training. In the experiment, our method achieved a similar level of mitigation of catastrophic forgetting while learning less data than PEWC.

Keywords Sampling network · Catastrophic forgetting · Dynamic mitigation

1 Introduction

Recently, a lot of researches has been conducted in the methodologies of learning a large amount of data due to the development of data collection capability and the improvement of computing power. In particular, the catastrophic forgetting [1] that occurs during transfer learning is a major obstacle. When multiple tasks exist, applying them to transfer learning results in a gradual degradation of performance on the datasets previously learned. A simple solution is to combine all the tasks into one set and relearn from scratch, but it is very inefficient in terms of cost. Elastic

D. Y. Hong · Y. Li · B.-S. Shin (✉)

Department of Computer Engineering, INHA University, Incheon, Republic of Korea

e-mail: bsshin@inha.ac.kr

D. Y. Hong

e-mail: 22181299@inha.edu

Y. Li

e-mail: leeyeon@inha.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_63

Weight Consolidation (EWC) [2] is designed to mitigate this, using a dynamic update technique using *Fisher Information*. However, EWC accepts all new tasks without screening so that it takes advantage of all images for learning. As a result, with many tasks, this technique is still ineffective in mitigating catastrophic forgetting. Predictive Elastic Weight Consolidation (PEWC) [3] further selects the data to be learned with a poor prediction by performing a prediction process for every new task. We propose a scheme for sampling the image to be learned dynamically, eliminating the sorting process of PEWC by learning extra network for only sampling separately. This results in a similar degree of mitigation of catastrophic forgetting while learning with less data than PEWC.

2 Sort-Free PEWC

PEWC is an improvement of EWC by introducing the sampling process. When PEWC encounters a new task, it does not proceed with learning immediately, but instead extracts a “*difficult*” image through a pre-prediction process. Pre-prediction is a new step proposed in PEWC. The new task is treated as a test set, and each image is scored with difficulty by L1-norm between the prediction and its actual annotation, and certain top data is sampled by sorting the images through the scores. PEWC defines an image with a high difficulty score as a difficult image. Catastrophic forgetting is caused by a loss of information due to a large number of updates. Therefore, PEWC uses less number of data by excluding images that could be accurately predicted in new tasks, which alleviates the forgetting.

However, the sampling of PEWC extracts difficult images based on a fixed criterion such as L1-norm of network prediction and the actual annotation, which requires not only the sorting process but also a new hyperparameter called sampling rate. Using a fixed sampling rate allows you to use a certain percentage of the image regardless of the overall difficulty of the task. This relearns easy images even when only a small amount of data is needed, thus hindering the mitigation of catastrophic forgetting. Sort-free PEWC (SF-PEWC) dynamically extracts difficult images by using a *sampling network* separately from the *learning network*. When a new task comes in, SF-PEWC uses the sampling network to determine whether each data applies to the learning network. As a result, the sampling network plays a role in extracting an image to be learned by the learning network, and the learning network plays the role of learning a task generated through a sampling network. We define the loss function of the sampling network as follows so that it dynamically selects the number of images to be used in each task (see Eq. (1)).

$$L(\theta_{sample}) = \|\theta_{sample}\|_2 + \|\theta_{learn}\|_1 + \|\theta_{sample}\|_1 \quad (1)$$

$$\text{where } \theta_{sample} : \text{parameter set of sampling network} \quad (2)$$

$$\theta_{learn} : \text{parameter set of learning network} \tag{3}$$

The sampling network uses L2-loss. The L1-regularizer is used together to prevent overfitting because it is a simple form that determines only whether the image is used in a learning network or not. Besides, the sampling network adds a learning network parameter set to the loss function to reflect the knowledge of the learning network. This idea is inspired by the teacher-student architecture [4], one of the knowledge distillation techniques. Unlike model compression, the main purpose of the teacher-student technique, the term is simplified to transfer knowledge between models.

The biggest difference between PEWC and SF-PEWC is the number of networks used. Previous PEWC conducted both a pre-prediction process and a learning process using only one network. The proposed technique uses a separate network for task sampling and does not include the sorting operation. Figure 1 is a schematic diagram of the predictive process of PEWC and SF-PEWC. The first column of Fig. 1 shows the image as a bar and the difficulty of the image as the height of the bar. As a new task comes in, PEWC predicts each image through a reference process and quantifies the difficulty of the image with L1-norm between prediction and annotation. It sorts all the images in the task according to the height of the bar, and samples a constant proportion of difficult images according to the given sample rate. On the other hand, the proposed scheme designs a sampling network with an output size of 1, and performs the reference process on the sampling network, not on the learning network. As a result, each image is immediately determined whether it is applied to the learning network. This eliminates the sorting process and allows for more dynamic task size adjustments because there is no limit to the number of images to be used.

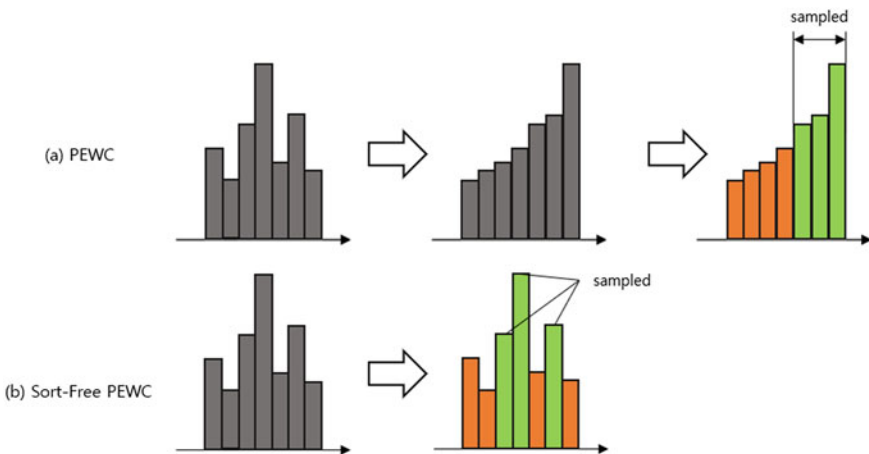


Fig. 1 Sampling process of PEWC and SF-PEWC. Schematic illustration of the difference between the pre-prediction processes of the PEWC and the SF-PEWC. Each bar represents an image and the height of a bar is the difficulty of it

3 Experimental Result

The experiment was performed using MNIST [5] and a multi-layer perceptron (MLP) [6]. Three experiments were designed to observe performance degradation trends for MNIST in EWC, PEWC, and SF-PEWC. Figure 2 plots the accuracy of the three cases, and Table 1 analyzes the accuracy measured during the learning process for the last task. Table 2 records the number of images for each task used in the learning network in each case, and Table 3 shows the time required for the learning process.

SF-PEWC showed no significant difference from PEWC in terms of learning time and accuracy while using fewer data through dynamic sampling (see Fig. 2b, c, Table 1). The proposed method and the PEWC showed less than 1% difference in the highest performance (Best), and the proposed method is 15% and 6% ahead in the worst and mean performance, respectively (See Table 1). In the graph, the proposed technique shows a more cohesive form than PEWC, indicating that the stability of learning did not decrease significantly (Fig. 2b and c).

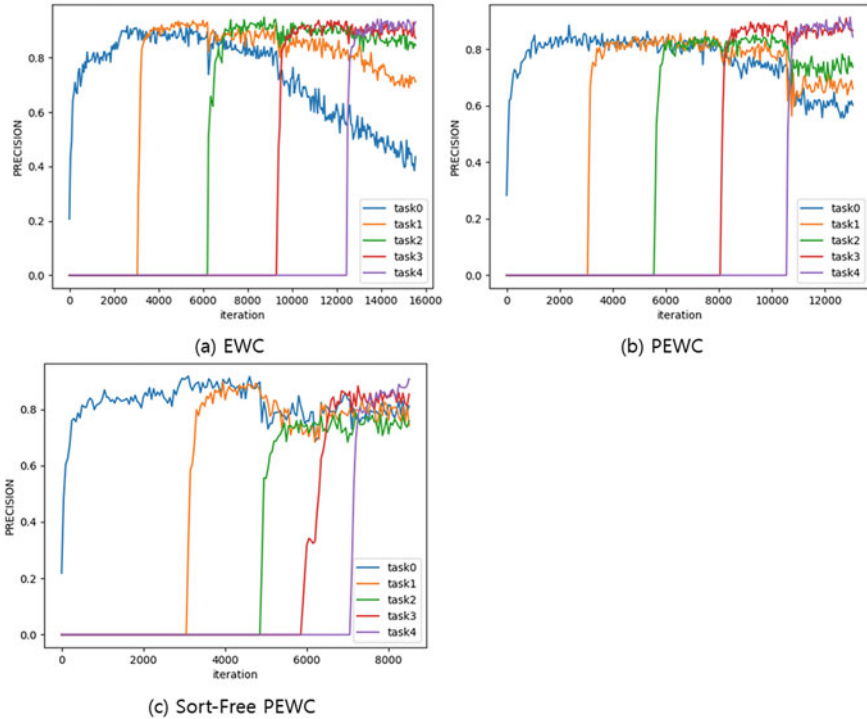


Fig. 2 Accuracies of EWC, PEWC, and SF-PEWC. This is the result of plotting the accuracy of all the tasks already learned for a specific iteration. At each point in time, the accuracy of the past task is checked to show the catastrophic forgetting

Table 1 Best, Worst, and Mean accuracies of each case in the last epochs. It shows the maximum value (Best), minimum value (Worst), and average accuracy (Mean) of all tasks measured during the final task learning. Best is the highest accuracy in the interval, and Worst is the lowest accuracy in the interval. Mean represents the largest value among average values of accuracy for all tasks at each point in the interval

	Best (%)	Worst (%)	Mean (%)
EWC	94.14	38.47	80.31
PEWC	91.40	55.66	76.48
Sf-PEWC	90.82	70.70	83.08

Table 2 # of images in each task. In each task, it represents the number of images used by the learning network to train. The unit is K

	T1	T2	T3	T4	T5	Total
EWC	50	50	50	50	50	250
PEWC	50	40	40	40	40	210
SF-PEWC	50	30.7	23	17.9	9.1	130.7

Table 3 Time for learning. In each case, it represents the total learning time of the learning network and the time required for each stage

	Pre-prediction (secs)	Train sampling network (secs)	Train learning network (secs)	Total (secs)
EWC	NA	NA	NA	152
PEWC	793	NA	152	945
SF-PEWC	50	703	91	844

The proposed technique is dynamic in the number of images to be extracted, and the number of images that make up each task is less than that of PEWC, resulting in about 38% fewer data used for total training (see Table 2). As a result, the learning time of the learning network was reduced by 40%, and the time taken to extract images was reduced by 94% due to the elimination of the alignment process (see Table 3). The total learning time was reduced by about 11%. The amount of computation has increased with the addition of a sampling network, but the total learning time has been reduced due to the benefit of a reduction in the number of data used.

4 Conclusion

Unlike PEWC, Sort-Free PEWC adds a separate network to perform the sampling process. PEWC uses sampling to alleviate the catastrophic forgetting that occurs during transition learning. However, PEWC's sampling method always uses only a

certain ratio of images given as hyperparameters. Therefore, the hyperparameter must be derived experimentally, and since all tasks use a fixed ratio, there is a possibility that many images are unnecessarily used for learning for a specific task. Sort-Free PEWC dynamically controls the number of images to be learned, removing the sorting process and a hyperparameter of the sample rate. This mitigates the worst oblivion with a similar level of PEWC while using fewer data.

References

1. McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In: Psychology of learning and motivation, vol 24. Academic Press, pp 109--165
2. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D, Hadsell R (2017) Overcoming catastrophic forgetting in neural networks. In: Proceedings of the national academy of sciences, vol 114, issue 13, pp 3521--3526
3. Dae Yong H, Yan L, Byeong-Seok S (2019) Predictive EWC: mitigating catastrophic forgetting of neural network through pre-prediction of learning data. J Ambient Intell HumIzed Comput 1--10
4. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y (2014) Fitnets: hints for thin deep nets. [arXiv:1412.6550](https://arxiv.org/abs/1412.6550)
5. LeCun YA, Cortes C, Burges CJ (1998) The MNIST database of handwritten digits. <https://yann.lecun.com/exdb/mnist/>
6. LeCun YA, Bottou L, Orr GB, Muller KR (2012) Efficient backprop, neural networks: tricks of the trade. Springer, Berlin, Heidelberg, pp 9--48

A Study on the Implementation of GRU Autoencoder Model for Detecting Insider Anomaly Behavior



Kyeong Geun Ryu and Deok Gyu Lee

Abstract With the development of high-tech technologies, the release of confidential information by industrial spy has also increased every year, causing damage to companies. If released, it could have a crucial impact on the nation and the national economy, and the outflow of core technologies continues increasing every year. This paper proposes machine learning to detect insider aberrations when there is a significant shortage of data about users in the company, such as the early part of the company, the new employees or the new security tool addition, to prevent the release of industrial data.

Keywords Machine learning · Security

1 Introduction

The number of crime leaking industrial secrets is increasing every year. As of 2013, the number of cases to be released overseas is increasing rapidly, in 2014, there are 472 cases of damage count and 50 trillion won in damages, equivalent to the annual sales of 4,700 small- and medium-sized companies. It is necessary to develop a pan-national response system to minimize such losses and to improve the company's

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00326, Development of smart port application platform by real-time tracking of logistics information based on blockchain).

K. G. Ryu · D. G. Lee (✉)

Department of Security Infomation, Seowon University, 377-3, Musimseo-ro, Seowon-gu, Choongbok, Cheongju-si, Republic of Korea

e-mail: deokgyulee@gmail.com

K. G. Ryu

e-mail: ryu8340@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_64

interest and awareness in establishing a security system for small and medium businesses. According to status of technology leakage in Industrial Confidential Protection Center, there are many cases of industrial secrets leakage in Korea's highly competitive precision machinery (34%), electricity and electronics (26%), and information and communication (14%), and it can be seen that the targeted technology of industrial spy is increasingly shifting and expanding from the technology of the IT sector of large enterprises to the precision machinery sector of small and medium companies [1, 2].

1.1 Purpose for Studies

In this study, we would like to study techniques in machine learning to detect security threats in a short period of time by detecting incidents of insider anomalies through rapid learning, even if the company's initial employees lack action data against unknown security threats.

2 Related Studies

2.1 Insider Threat Detection

It is not easy to accurately categorize and preemptively block all threats from insiders in network, such as mistakes by normal users and unauthorized access by disgruntled users, and thus need the means to effectively detect their activities from within. Given that the average amount of time it takes to discover data breaches is 191 days, the average days to suppress them is 66 days, and the average cost per breach is \$3.62 million, the effect of detecting suspicious user activity early would be enormous [3]. In order to analyze an insider's threat behavior, the user sends a log based on the user's behavior (file movement, network login, etc.) to machine learning, which determines the threat behavior. For method of judgment the threat behavior, suggest to use combining GRU (Gated Current Unit) and the Autoencoder.

2.2 Specific Behavior Detection

The most common method is to have an administrator on the DB Server to recognize specific activities (such as file movement, copying, and network login) to manage suspicious behavior as an administrator. The use of this method allows for precise inspection, but has a problem. If users modify or move multiple files in a short time, active response is not possible.

2.3 Detection by Learning Machine Learning

The issue of insider threats has become a major issue that has been studied steadily since the past, and is increasingly being studied. With the recent development of machine learning, there are several preceding studies that have applied it to the field of insider threat detection. The most representative method is the model that is detected using the Hidden Markov Model (HMM). HMM has been widely used to recognize sequential data, such as voice data, as a model that can be well applied to the problem of dealing with data that implies a set of sequential characteristics [4].

2.4 Algorithm

Machine learning technology is largely divided into a Supervised Learning and Unsupervised Learning. Supervised Learning looks for labels by comparing them to the model that generated the unlabeled data the same Support Vector Machine as Classification and the same Linear Regression as Regression after generate model that learned by data label in it. There is clustering, such as K-means clustering, which is a way to find labels after classifying them without labels [5]. In this paper, we are going to combine the GRU model in case the insider aberration data used is small and the AutoEncoder, a Unsupervised Learning.

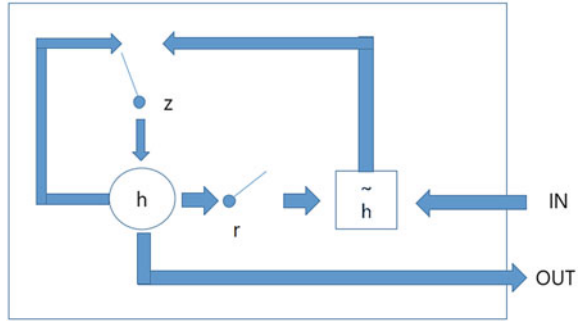
2.5 LSTM (Long Short Term Meometry)

RNN significantly reduces its ability to learn reverse waves when the distance between the relevant information and the point at which it is used is far. Designed to overcome this problem is LSTM (Long Short Term Meometry). The problem was solved by adding a cell-state to the Hidden State of RNN by recursively obtaining it. LSTM network operates using cells consisting of forget gate (FG), input gate (IG), cell gate (CG), and output gate (OG). These memory cells can store information at any time. The three gates control the flow of information into and out of memory cells in neurons. Each gate in the LSTM receives the same input as the input neuron, each with an activation function [6].

2.6 GRU (Gated Recurrent Unit)

The Gated Current Unit (GRU) is one of the LSTM variant models that handled LSTM's structure more simply with the model announced in 2014. The structure of the GRU, like LSTM, is the same as using gates to control the amount of information,

Fig. 1 GRU model framework

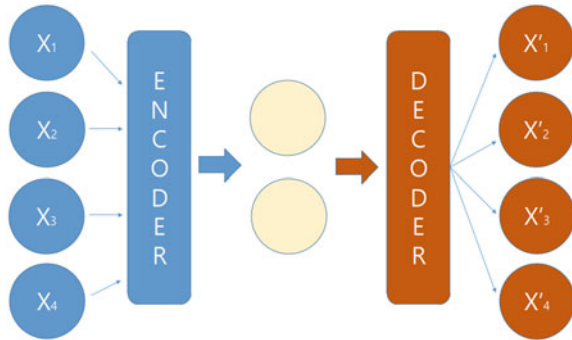


but it can be seen that there are differences in the way gates are controlled. The GRU integrated LSTM's forget gate and input gate into the update gate, and integrated the cell state and the hidden state into one. It is also faster to learn because of its smaller weighting structure than LSTM, but has almost the same performance as LSTM [7]. The GRU has fewer parameters, making it easier to add or modify inputs. It also takes shorter learning times and is less frequent hyperconformity phenomenon than LSTM. Less data may be learned than LSTM, but in other words, with enough data, LSTM's good modeling power may show better results. As a result, in this paper, GRU will be used as an early model without data (Fig. 1).

2.7 AutoEncoder

The Autoencoder is an Artificial Neural Networks (ANNs) studied to compress image data. It is a representative model for unsupervised learning where AI learns on its own. The structure of the Autoencoder produces an invisible output value z of x received as input through the encoder, and produces the final output value x' with z as the input of the decoder. Having the final output value x' and the input value x have the same value after this process is called the Autoencoder. The purpose of the Autoencoder is to create a model that contains compressed expressions of input values while repeating the encoding and decoding processes in the course of learning. Using these features, Autoencoder is being used in many areas, including abnormal detection and data generation model learning [8]. Data for user behavior analysis has time-series data, which is very large. It attempts to implement an insider aberration detection model by combining models of AutoEncoder and GRU that can compress data and have a unsupervised learning process (Fig. 2).

Fig. 2 AutoEncoder structure



3 Proposal Method

3.1 Process of User Behavior Analysis

From each employee PC, the user’s behavior data is continuously transferred to the collection server, using this program. The server sends this data to Machine Learning to determine if the behavior is abnormal and sends the logs that it determines back to the collection server. If abnormal behavior appears here, block the PC.

3.1.1 Machine Learning Normal Behavior

In this paper, in order to detect abnormal behavior of users, the user will learn the normal behavior of the user and determine whether the user is a normal behavior or not. In this case, there is a standard for determining. There will be a sequence of daily tasks for each user, and they will be carried out in a similar. In addition, each individual has a certain pattern and will learn this data and judge it as normal behavior. Besides, there are two other ways to divide an insider’s threat into two behaviors. First, behavior that deviates greatly from normal work, such as network access, copying to unnecessary files, and the second USB overuse connections, and psychological factors, which are the weak psychological state and hostile behavior of the user concerned. Based on this, they learn normal behavior in machine learning.

3.1.2 Configuring User Behavior Patterns

Process behavior patterns are constructed according to a list of patterns produced through process analysis, and user behavior patterns are constructed by established behavior pattern assigning additional points depending on the type of system. The pattern is redefined by assuming normal operating conditions for the derived pattern and weights are applied by security element according to the system’s configuration

characteristics. For additional points, the system is constructed to reflect the characteristics of the system, such as type of operation, repeatability, frequency, etc. and the weight calculated is applied by element to calculate the process function score in normal circumstances.

3.1.3 Pre-processing Normal Behavior Data

Data analysis is necessary to study the machine. Data analysis can eliminate unnecessary information and maximize the effects of machine learning. As a collection of data, data containing user logon/logoff records, web activities, file access activities, e-mail usage, device use and user position, department, work period, participation projects, and job satisfaction will be collected in a chronological order, along with other psychological propensity indicators for each employee.

3.1.4 GRU-Autoencoder

This paper uses GRU Autoencoder. This model consists of three GRU Encoder and GRU Decoder, which sequentially reads the input data and thereby learns 'h', a size representation that is fixed in a hidden state. This expression is passed to the GRU Decoder with the cell state and output state of LSTM. The passed values maintain sequential and user attribute information for user actions. By taking 'h' as input, GRU Decoder generates an output sequence sequentially. In this process, connect three GRUs in. Since errors can occur when reconstructing the result values, a loss function called Cross Entropy is used to optimize the reconstruction to minimize the occurrence of errors. It also uses a technology called Drop-out to prevent over-conformity. Over-conformity can cause performance to deteriorate if there is a lot of information in machine learning. At this time, it is Drop-out that properly disconnects information to help you learn only some information. In this way, the output sequence is obtained by learning by minimizing errors (Fig. 3).

3.2 Machine Learning Abnormal Behavior Detection

When you have finished learning, each user has an output sequence for normal behavior. Use this to calculate the weight. In addition, if this weight is set to a threshold, and subsequently continues to receive logs from the user's computer, and then compare them to normal behavior to users who had the same or similar patterns as normal patterns, the weight does not exceed the threshold and is considered normal behavior, and if abnormal patterns, the weight is significantly different from the normal pattern and thus detects abnormal behavior.

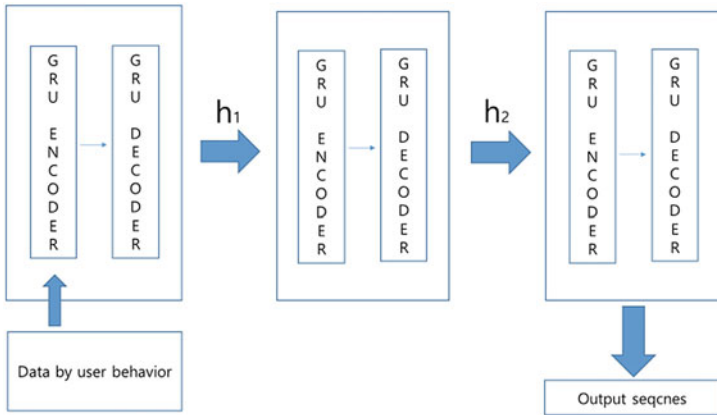


Fig. 3 GRU-Autoencoder model

4 Analysis

Figure 4 presents a comparison between the existing RNN and LSTM user behavior analysis and the system proposed in this paper. In terms of data dependency, traditional RNN and LSTM have very high data dependence on user analysis. In contrast, GRU is a proposed model that reduces data dependence at LSTM, so RNN and LSTM learn 300 nodes, and GRU learn 12 nodes with similar accuracy [9]. As a result, abnormal behavior can be inferred faster than previous methods, even if new employees enter the company or there is a lack of data such as changes to security equipment. In terms of accuracy, RNN is long-term dependent in the way that the hidden node circulates, and if the distance between the points where the information is used is far from the point where the learning ability is significantly reduced and performance is degraded. As an LSTM model designed to solve these problems, GRU is an algorithm created by simplifying the LSTM model. In the GRU model,

	RNN	LSTM	GRU
Overfit frequency	●	●	●
accuracy	▲	●	●
Overfit frequency	●	▲	X

Fig. 4 Evaluation result

high accuracy can be achieved with less data and shorter learning times. LSTM has lower accuracy than GRU in small data, but has shown excellent results when there is a sufficient number of data. In terms of the frequency of overconformities, RNN is prone to overconformity as there are no forgettable gates in the way that the hidden node rotates. LSTM has reduced the overconformity ratio as it is an architecture that adds memory cells to RNN. In here, GRU integrates cell state and hidden state to further reduce load on memory and reduce the frequency of overconformities. This paper also used drop-out technology to further reduce overconformity. Based on the analysis, analysis of user behavior using LSTM and GRU has superior performance than analysis of user behavior using RNN. Where LSTM showed higher accuracy when it had a lot of data and GRU had lower accuracy than LSTM when it learned a lot of data, but GRU showed higher accuracy in a short period of time with less data.

5 Conclusion

With the development of high-tech technologies day by day, there has also been an increase in confidential information leaks by industrial spy every year, resulting in huge damage to companies. Thus, even in-house, abnormal behavior detection tools make it difficult to accurately determine which employees are internal leakers due to lack of data when there is no data about users in the initial use and when new employees join the company, when data are added to the security tools as a result of changes. To improve this problem, this paper proposes a GRU model that uses the GRU model for analysis machine learning to detect abnormal behavior within a relatively short period of time, thereby preventing leakage of confidential industrial data.

References

1. Lee SO (2019) A study on the effective method for the prevention of industrial secrets leakage. *Chung_Ang Law Rev* 21(1):39–80
2. Jeong JH (2015) USB serial number. <https://intro0517.tistory.com/107>
3. Lee JD (2007) SSDT hooking, [index-of.co.uk/Reverse-Engineering/Hide process](http://index-of.co.uk/Reverse-Engineering/Hide%20process)
4. Kim WK, Soh WY, Sung K (2009) Study on the API hooking method based on the windows. *J Korea Navig Inst* 13
5. Lee YW (2018) Insider anomaly detection method using machine learning. *Korea Institute of Information Scientists and Engineering*, pp 971–973
6. Joe CH (2018) Research on an efficient Insider threat detection based on LSTM Autoencoder using user attribute information. *Sungkyunkwan University Doctor's Thesis*
7. Chung J, Gulcehre C, Cho KH, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modelling. In: *NIPS 2014 deep learning and representation learning workshop*

8. Ha DW (2018) A study on a machine learning model for detecting insider anomaly behaviour.,Myongji University Master's Thesis
9. Kim HH (2017) Forecasting time-series data using LSTM/GRU recurrent neural networks. Korea National Open University Master's Thesis

Blockchain Based Authentication Method for ThingsBoard



Sung Il Jang, Ji Yong Kim, Alisher Iskakov, M. Fatih Demirci,
Kok Seng Wong, Young Jong Kim, and Myung Ho Kim

Abstract The X.509 certificate is generally used for a device authentication of the internet of things platforms. X.509 certificates use the device's public key based communication for device integrity. However, this process incurs a large overhead. Access tokens, on the other hand, have less overhead but have problems with integrity. In the paper, we propose a new authentication scheme based on blockchain. We conducted several experiments to prove that proposed method have less overhead than X.509 certificates, and that the authentication scheme works appropriately.

Keywords Internet of things · Internet of things platform · Authentication · Blockchain · ThingsBoard · Hyperledger fabric

S. I. Jang · Y. J. Kim · M. H. Kim (✉)
Department of Software, Soong-Sil University, Seoul, Korea
e-mail: kmh@ssu.ac.kr

S. I. Jang
e-mail: sungil@soongsil.ac.kr

Y. J. Kim
e-mail: youngjong@ssu.ac.kr

J. Y. Kim
Department Of Software Convergence, Soong-Sil University, Seoul, Korea

A. Iskakov · M. Fatih Demirci · K. S. Wong
Department of Computer Science, Nazarbayev University, Nur-Sultan City, Kazakhstan
e-mail: alisher.iskakov@nu.edu.kz

M. Fatih Demirci
e-mail: muhammed.demirci@nu.edu.kz

K. S. Wong
e-mail: kokseng.wong@nu.edu.kz

1 Introduction

Internet of Things (IoT) platforms are the software platforms optimized for IoT data collection and management to establish the traffic of machine-to-machine (M2M) circuits that has recently been widely implemented [1]. ThingsBoard is one of the platforms developed to collect and manage the data of interconnected devices by using various protocols such as the MQTT, HTTP and the CoAP through the IoT gateway. It is aimed to ensure compatibility with proprietary solutions such as Amazon web services (AWS) IoT [2, 3].

As a device authentication method, ThingsBoard employs access tokens and X.509 certificates, which are referred as one-way secure sockets layer (SSL) and two-way SSL, respectively. One-way SSL is the encrypted communication using the public key of the server, and two-way SSL is that one using the public key of both the server and the device [4]. Each authentication method has its pros and cons. The X.509 certificate provides a high level of security owing to encrypted communication based on the public key between the server and the device. However, encryption and decryption operations on the device have an impact on the battery usage. Moreover, it takes longer time to do authentication comparing with the method using access tokens. Using the access token has a disadvantage that the data can be easily leaked from unencrypted networks. However, this method does not significantly effect on the battery usage and has less network overhead comparing with the X.509 certificate, as there are no additional operations needed during the authentication process. But it is difficult for the access token to guarantee the integrity of devices. Taking into account the characteristics of the IoT environment, in this paper, we solve this problem by sharing the access token through the blockchain and proposing a new authentication scheme based on the shared access token.

2 ThingsBoard

ThingsBoard uses the access token and the X.509 certificate for authentication of devices. The identity of each device is ensured by the token value and the public key. The architecture of ThingsBoard is shown in Fig. 1 [5]. ThingsBoard recognizes the gateway as a device. The gateway forwards messages from the device to the server. If the X.509 certificate is used, the identity of the device is proved by the public key, even if the gateway forwards messages in the middle of the process. However, in the case of using the access token, when the gateway forwards the message of the device, the IoT integration middleware (IoTIM) can verify only the token value of the gateway. And the gateway cannot verify the device's token value, because it does not have access to the central database. In the IoT environment it is difficult to provide the gateway within granted access to the central database due to the large number of gateways.

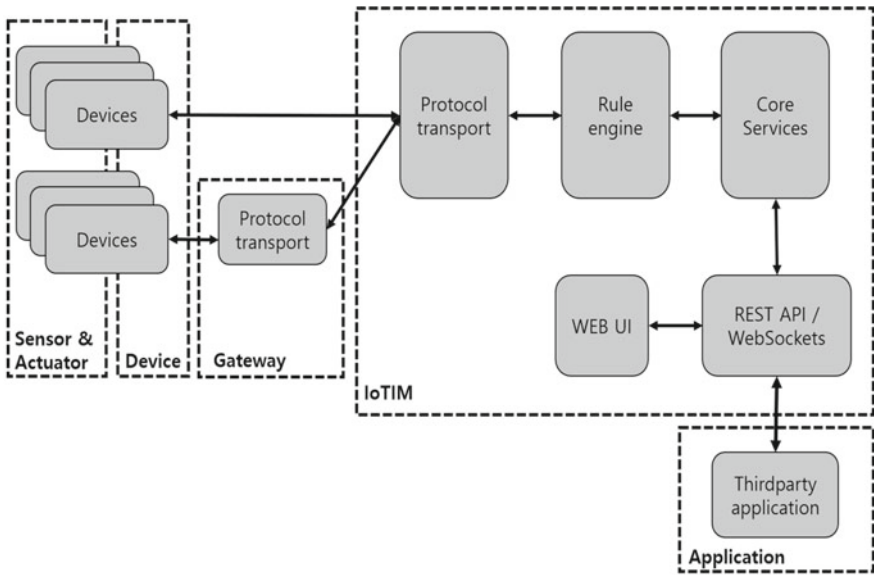


Fig. 1 ThingsBoard architecture based on IoT reference architecture

3 Blockchain

The blockchain ensures the data by managing transactions on a block basis and sharing the chains of consecutive blocks. In the process of sharing blocks, the blockchain uses a consensus algorithm. According to this algorithm, the blockchain can be classified into the permissionless blockchain and the permissioned blockchain [6]. In this paper, we used the permissioned blockchain that was relatively fast and effective in managing the blockchain verification node, as it distributed and stored access tokens through the blockchain and allowed authenticating the device transmitting the data. The permissioned blockchain can be implemented by means of various platforms corresponding to a particular consensus algorithm. After considering different options, we conducted the experiments using Hyperledger Fabric that employs the consensus algorithm based on practical Byzantine fault tolerance (PBFT) [7]. Hyperledger Fabric establishes the PBFT consensus algorithm using the process shown in Fig. 2 [8].

Hyperledger fabric is suitable for storing access tokens, as it can guarantee the finality of transactions, unlike the permissionless blockchain that uses the Proof of Work (PoW) algorithm.

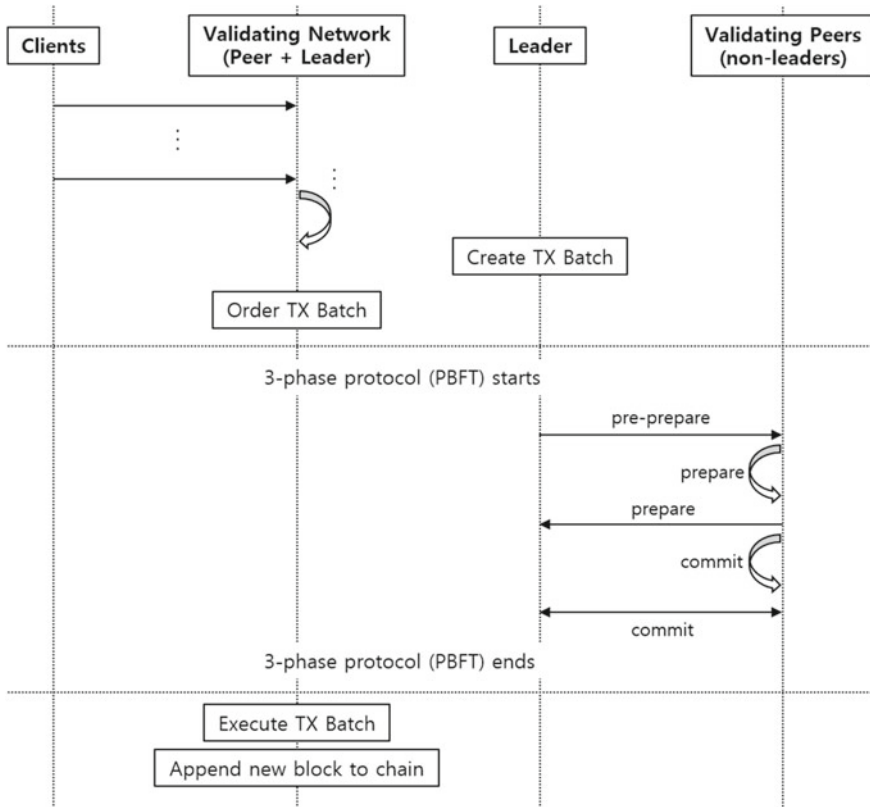


Fig. 2 Sequence of PBFT consensus in Hyperledger Fabric

4 Proposed Method

We propose the architecture presented in Fig. 3 to share access tokens by using the blockchain in the conventional ThingsBoard architecture. The existing IoTIM stores the access token value by using the database, however, the proposed method stores it in the blockchain also. This enables the gateway to verify the token value of the device.

Figure 4 shows the authentication scheme of the proposed method. According to this scheme, the gateway logs into IoTIM using the access token as same to the conventional method. After this, when the gateway receives the data from the device, it forwards the data to IoTIM by executing the onDeviceData function. The gateway performs device authentication using the token value delivered to the device through the chaincode (known as Smart Contract) named a device validator. The device with the correct token transmits its data to IoTIM by executing the onDeviceConnect function.

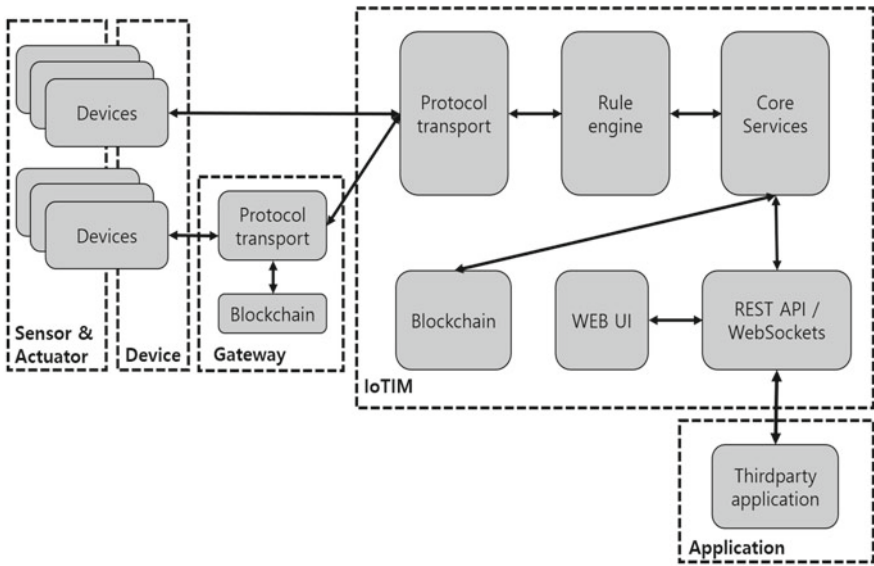


Fig. 3 Proposed ThingsBoard architecture using blockchain

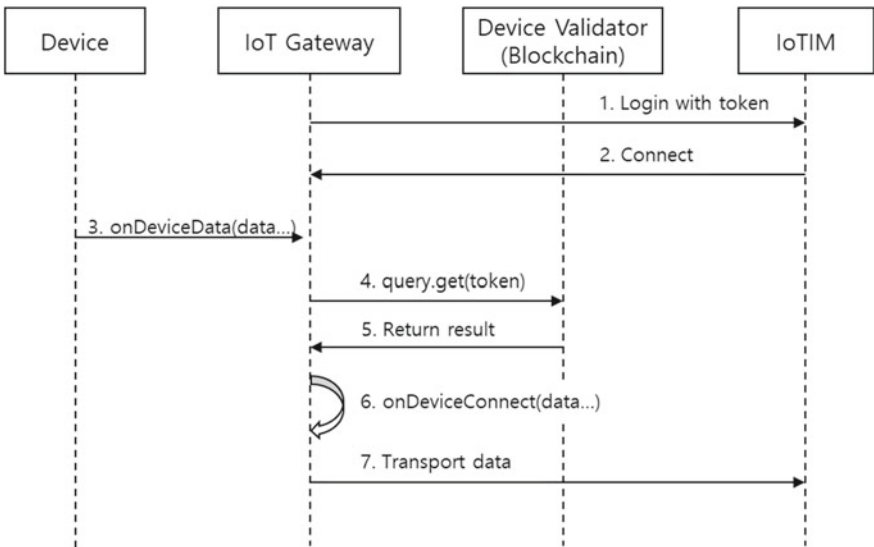


Fig. 4 Sequence of steps in the proposed authentication scheme

The device validator can be executed locally at each gateway, because it is done through the blockchain. Therefore, the proposed gateway is suitable for the IoT environment, as it does not cause a network overhead, whenever the device is authenticated.

5 Implementation and Testing

5.1 Overview

In this paper, we used ThingsBoard and ThingsBoard IoT Gateway as the IoT platform, and Mosquitto as the MQTT broker to implement the proposed method. We employed Hyperledger Fabric v0.6 and updated the Go language to the version 1.11 to implement the chaincode.

In the first experiment, we compare the message processing time between two authentication methods within ThingsBoard. In the second experiment, we confirm the results by sending messages to the gateway applying each method to verify the difference between the existing method and the proposed method, when authentication is performed at the gateway. The format of the message is described in Table 1, and two messages are sent using each method: the first message forwards the type of the device to the attribute value, and the second message forwards the token value. This is summarized in Table 2.

5.2 Testing and Evaluation

In the first experiment, the device send a message 20 times using each authentication method. The obtained results are shown in Fig. 5. The proposed method have less overhead than X.509 certificate approximately 23%.

The results of the second experiment are shown in the tables below. Analyzing the results provided in Tables 3 and 4, we can conclude that the messages corresponding to Case 1 and 2 sent from the existing gateway are all forwarded to IoTIM correctly. The existing method does not have an authentication scheme through the gateway, therefore, it forwards the data to IoTIM without being affected by any message values.

Table 1 Message format of second experiment

Name	Role
SerialNumber	DeviceName and identifier
Model	DeviceType and attributes
Temperature	Telemetry data

Table 2 Message of second experiment

Case No	Message	Destination
Case 1	{“serialNumber”: “Device005”, “model”: “T1000”, “temperature”: 33.0}	Existing method based gateway
Case 2	{“serialNumber”: “Device005”, “model”: “NL4R74HgDaIXbZhEkyYi”, “temperature”: 34.0}	Existing method based gateway
Case 3	{“serialNumber”: “Device005”, “model”: “T1000”, “temperature”: 35.0}	Proposal method based gateway
Case 4	{“serialNumber”: “Device005”, “model”: “NL4R74HgDaIXbZhEkyYi”, “temperature”: 36.0}	Proposal method based gateway

Fig. 5 Processing time for the two considered device authentication methods in ThingsBoard

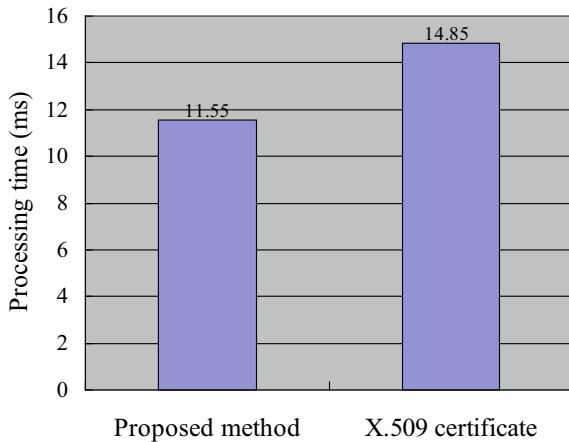


Table 3 Result of Case 1

Log message
19:56:48,769 INFO [Device005] Device Connected!
19:56:49,341 INFO [Device005] [T1000] [10] Device connect event is reported to Thingsboard!

Table 4 Result of Case 2

Log message
19:58:27,851 INFO [Device005] Device Connected!
19:58:28,353 INFO [Device005] [NL4R74HgDaIXbZhEkyYi] [6] Device connect event is reported to Thingsboard!

Table 5 Result of case 3

Log message
20:02:34,994 INFO [T1000] Failed to report device connection (TokenError, onDeviceData)!

Table 6 Result of case 4

Log message
20:00:35,855 INFO [Device005] Device Connected!
20:00:36,420 INFO [Device005] [NL4R74HgDaIXbZhEkyYi] [14] Device connect event is reported to Thingsboard!

Analyzing the results provided in Tables 5 and 6, it is evident that the proposed gateway processes the forwarded value as an attribute of the device in a form of an access token value and then, performs authentication accordingly. As shown in Table 5, TokenError occurred due to the fact that “T1000,” not the token value, was forwarded, while Table 6 shows that the data were sent to IoTIM, as the correct token value was forwarded.

6 Conclusions

In this paper, we analyzed the process of the authentication within the IoT platform called ThingsBoard. We outlined a problem that the gateway could reduce the integrity of the device within the IoT platform while forwarding messages of the device, and proposed an improved authentication scheme using blockchain to solve this problem.

In the first experiment, we compared the device’s message processing time between the proposed method and the X.509 certificate. Through this, we estimated the overhead difference between the two considered authentication methods and confirmed the necessity of employing the proposed method in the IoT environment. In the second experiment, we confirmed that the gateway was able to authenticate the device using the proposed authentication scheme. Therefore, in this paper, we solved the problem of the access token (i.e., the integrity problem) using the blockchain and consequently, we reduced authentication overhead than X.509 certificate.

Acknowledgements This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01419) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

1. Cisco: cisco visual networking index: global mobile data traffic forecast update, 2017–2022. Cisco public (2019)
2. Alexandr K, Marco J, Edoardo P (2015) Designing a smart city internet of things platform with microservice architecture. In: The 3rd international conference on future internet of things and cloud, pp 25–30
3. Eduardo NB, Willian PDA (2018) IoT centralization and management applying ThingsBoard platform. Project report, Häme University of Applied Sciences
4. ThingsBoard device authentication options. <https://thingsboard.io/docs/user-guide/device-credentials/>
5. Guth J, Breitenbücher U, Falkenthal M, Fremantle P, Kopp O, Leymann F, Reinfurt L (2018) A detailed analysis of IoT platform architectures: concepts, similarities, and differences. In: Internet of Everything. Springer Nature, New York, pp 81–101
6. Zibin Z, Shaoan X, Hongning D, Xiangping C, Huaimin W (2017) An overview of blockchain technology: architecture, consensus, and future trends. In: IEEE international congress on big data (BigData Congress), pp 557–564
7. Christian C (2016) Architecture of the hyperledger blockchain fabric. IBM Research
8. Harish S, Jos é MM, Xiaolin C, Kishor ST, Andy R (2017) The physiology of the grid: performance modeling of PBFT consensus process for permissioned blockchain network (Hyperledger Fabric). In: IEEE 36th symposium on reliable distributed systems (SRDS), pp 253–255

Secure Management of Patient Medical Data Using QR Code and CP-ABE



Su-Mee Moon, Beakcheol Jang, Hoon Yoo, and Jong Wook Kim

Abstract QR codes are widely used in various services provided by agencies such as banks, credit card companies, and airlines. Among these, QR code based data management for the medical data of hospital patients is a representative area of QR code applications. As medical data usually contain sensitive patient information, preventing their leakage is a serious concern. Therefore, in this study, we present a method for the secure management of patient medical data by leveraging QR code and ciphertext-policy attribute-based encryption (CP-ABE). Especially, the proposed method can be used to securely store sensitive patient medical data in a QR code, which allows access only to authorized persons.

1 Introduction

The advent of two-dimensional codes with patterns such as squares and hexagons has widened the range of information storage. Specifically, the quick response (QR) code, which was invented by Denso Wave Inc., Japan in 1994, has been actively used in more than 20 types of two-dimensional codes [1]. The QR code can store up to 7,089 decimal digits, which is much higher compared to the barcodes that can

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korean Government (MSIP) (Development of Integraphy Content Generation Technique for N-Dimensional Barcode Application) under Grant 2017-0-00515.

S.-M. Moon · B. Jang · H. Yoo · J. W. Kim (✉)
Department of Computer Science, Sangmyung University, Seoul, Korea
e-mail: jkim@smu.ac.kr

S.-M. Moon
e-mail: sumeedi@naver.com

B. Jang
e-mail: bjang@smu.ac.kr

H. Yoo
e-mail: hunie@smu.ac.kr

store only up to 20 characters [2]. Moreover, it is used in various fields, owing to its advantages that include recognizability from all directions through the various symbols placed in it, information recovery ability, and security function.

Recognition of the QR code uses the principle of light reflection, similar to the barcodes; 0 and 1, which are represented by black and white bars, respectively, are scanned by a camera and converted into digital signals. As mobile devices with built-in cameras have become commonplace in recent times, QR codes are being considered as one of the leading data repositories, owing to their speed, portability, and security. Thus, QR codes are used in various fields, the most representative being the medical field. For example, patient ID cards containing medical information such as past treatment records or allergies are currently used in many medical institutions [3].

However, although the QR codes offer the advantage of easy storage and access to information, they are associated with privacy issues: for example, a QR code containing patient information attached to a hospital room can be easily scanned and misused. As medical data include sensitive information about individuals such as drug side effects and allergies, as well as general information such as name and age, it is essential to address the privacy issues while applying the QR code in the medical field.

Therefore, in this study, we present a method to preserve the medical data stored in QR codes: the patient medical data is first encrypted using ciphertext-policy attribute-based encryption (CP-ABE), which is a kind of attribute-based encryption (ABE) technology that defines access rights through attributes. Then, the encrypted patient medical data is stored in a QR code. Since the patient medical data is encrypted by using CP-ABE, the encrypted data can be decrypted differently, based on the accessor's attributes. The hierarchical structure of the hospital or access authority determines the information that can be retrieved, thereby protecting sensitive patient information from attackers.

2 Background: Ciphertext-Policy Attribute-Based Encryption

Attribute-based encryption is a public-key encryption method wherein a user is identified through a set of attributes such as the name and title. Owing to their wide expressiveness, attributes are actively used in cloud storage and other similar applications [4]. Conventional public-key cryptography systems use a public key that is known to everyone and a private key that only the recipient of the message knows. For example, when there are several users, a message is encrypted using the public key, and the message is decrypted using the private key [5]. In contrast, in ABE, the public key is the same for all users, while the private key associated with an attribute set, which is different for each user. Therefore, the ABE technology is effective in terms of storage space and encryption time, because only one encryption file is generated.

The ABE is usually categorized into key-policy attribute-based encryption (KP-ABE) and CP-ABE. While the former encrypts a file with a set of attributes and generates a user's private key through policy, CP-ABE encrypts files with a policy and generates a user's private key with a set of attributes. The disadvantage of KP-ABE is that it is not intuitive because the relationship between the properties cannot be specified. Moreover, it is not possible to know the attribute required to decrypt the ciphertext for each user [6]. On the other hand, CP-ABE encrypts files such that it allows for an intuitive definition of the data accessors. Thus, it is easy to assign and delete permissions by defining each user with a set of attributes, thereby avoiding authorization creep.

The CP-ABE used in this study encrypts medical information by defining policy with an attribute set and logical operators. The policy is stored in a built-in form while encrypting medical information: the user can decrypt the encrypted information with an attribute that satisfies the policy [7]. The CP-ABE consists of the following four basic algorithms [6]:

1. *Setup*: Algorithm that returns the public key PK and master key MK
2. *Encrypt* (PK, M, W): Algorithm to return CT , an encrypted message by inputting PK , message M , and access structure W
3. *KeyGen* (MK, L): Algorithm to return secret key SK by inputting MK and attribute set L
4. *Decrypt* (PK, CT, SK): Algorithm that inputs PK , encrypted message CT , and SK , and returns decrypted message M if the attribute satisfies the access structure.

3 CP-ABE-Based Medical Data Management Using QR Code

In this section, we describe the proposed method for securely storing and accessing patient medical data using a QR code. The proposed approach relies on CP-ABE to support multilevel privacy protection mechanisms depending on the trust level between the patients and authorized data users. Figure 1 shows the system architecture of the proposed approach, which mainly consist of two parts: the storing of medical data in a QR code via CP-ABE and accessing the data stored in it. The following is a detailed explanation of the proposed approach.

3.1 Storing Medical Data in a QR Code via CP-ABE

Figure 2a shows the user interface developed in this study for entering patient information including the name, date of birth, height, weight, and blood type, which have a single value, and drug allergy and chronic disease, which can have multiple values. On entering the patient information through the user interface and clicking on the

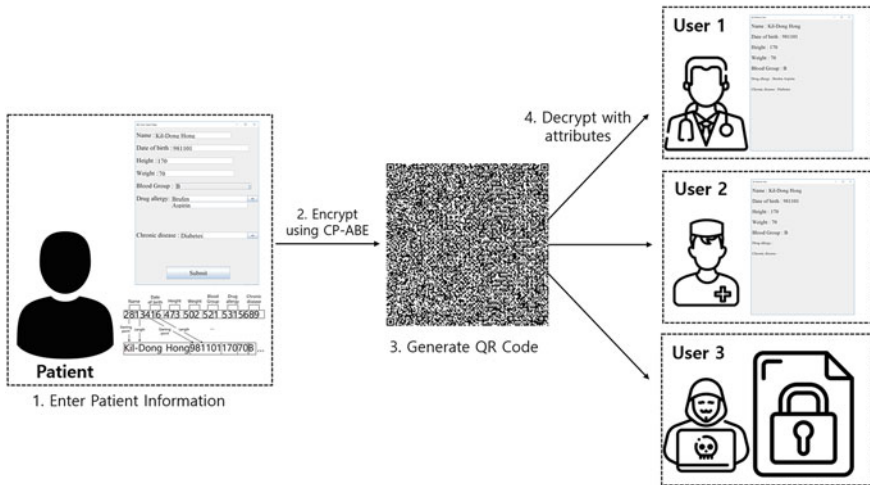


Fig. 1 The system architecture of the proposed approach

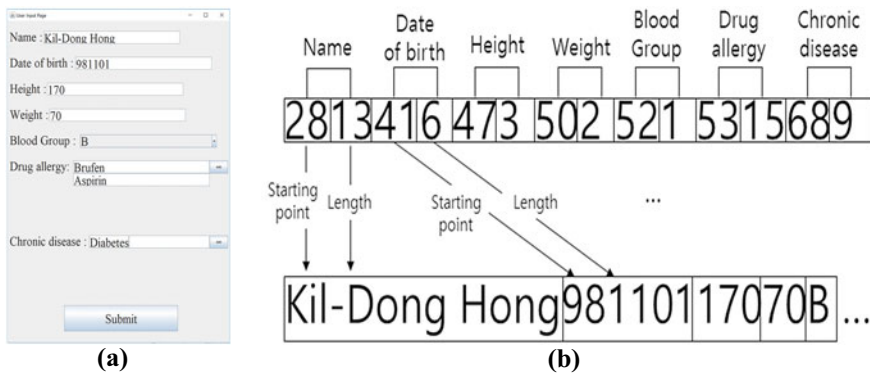


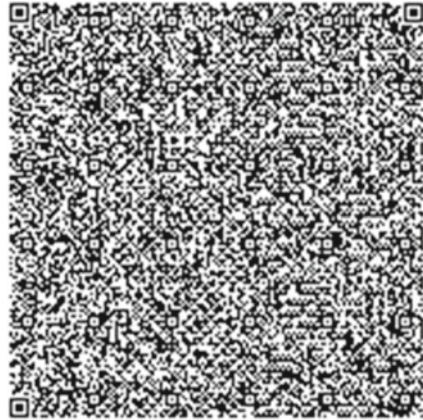
Fig. 2 **a** User interface developed for entering patient data, and **b** the corresponding plaintext, which consists of a header and data parts

submit button, a corresponding plaintext that consists of two parts (namely a header part and a data part) is generated.

Figure 2b shows the code generated based on the information provided in Fig. 2a, which includes the header part, containing a list of (*offset, length*) that indicate the position of the actual data stored in the plaintext: here, *offset* denotes the beginning of the corresponding data within the plaintext, and *length* represents the length of data. For example, in Fig. 2b, the name of the patient is stored from the 28th location, and its length corresponds to 13.

In the next step, the plaintext is encrypted using CP-ABE. In this study, the patient data is classified into two groups according to the degree of sensitivity of the

Fig. 3 Sample QR code generated by the proposed method



information: The low-sensitive data group includes the name, date of birth, height, and weight, while the high-sensitive data group contains details of drug allergies and chronic diseases. Each data in these two groups are encrypted via CP-ABE. The access rights of user groups in CP-ABE can be set differently, as required. In this study, we assume a scenario wherein three different user groups attempt to access patient medical data: The first is the doctor group, the second is the nurse group, and the third is the attacker group. The access condition for the low-sensitive data (i.e., name, date of birth, height, and weight) is ‘doctor’ or ‘nurse’. The access condition for the high-sensitive data (i.e., drug allergies and chronic diseases) is ‘doctor’. Figure 3 shows the encrypted medical data stored in a QR code, which was generated using ZXing, an open source provided by Google.

3.2 Accessing Medical Data Stored in a QR Code

Figure 4 presents an example scenario wherein users from different groups attempt to access the medical data stored in the QR codes. USER 1 belongs to the doctor group, which has the highest data access rights. Thus, USER 1 can decrypt and access both the high- and the low-sensitive patient medical data. USER 2 belongs to the nurse group whose access rights are lower than those of the doctor group. Thus, USER 2 can only decrypt and access the low-sensitive patient medical data. Finally, USER 3, belongs to a group of attackers and therefore, cannot access any medical data stored in the QR code.

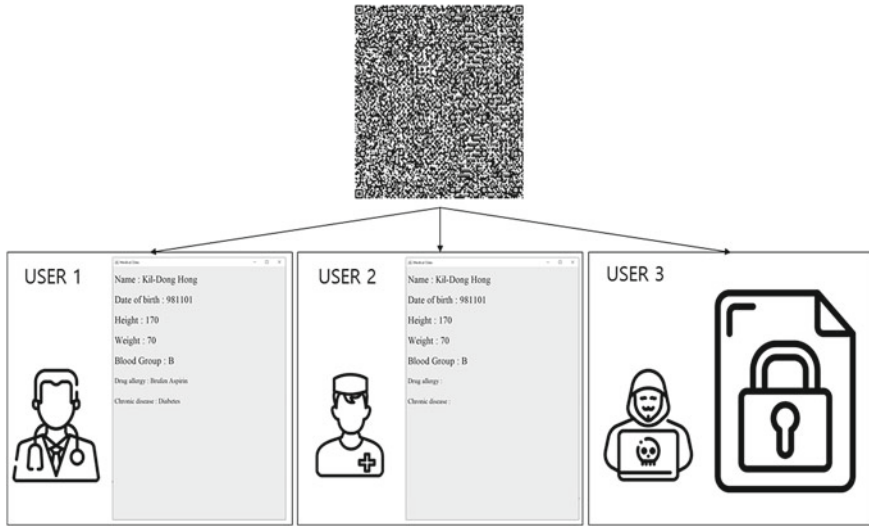


Fig. 4 An example scenario wherein users from different groups attempt to access patient medical data stored in a QR code

4 Conclusion

In this study, we have presented a secure method for storing and accessing patient's medical information using the QR code and CP-ABE. The proposed method first encrypts sensitive patient medical data using CP-ABE, and then stores the encrypted data in a QR code. Therefore, only authorized persons can decrypt the encrypted data, and access sensitive medical information. Furthermore, by using CP-ABE, the presented method can support multilevel privacy protection mechanisms depending on the trust level between the patients and authorized data users.

References

1. Melgar MEV, Farias MC (2019) High density two-dimensional color code. *Multimed Tools Appl* 78(2):1949–1970
2. Grillo A, Lentini A, Querini M, Italiano GF (2010) High capacity colored two dimensional codes. In: *Proceedings of the international multi conference on computer science and information technology*. IEEE, pp 709–716
3. Uzun V (2016) QR-code based hospital systems for healthcare in Turkey. In: *2016 IEEE 40th annual computer software and applications conference (COMPSAC)*, vol 2. IEEE, pp 71–76
4. Green M, Hohenberger S, Waters B (2011). Outsourcing the decryption of abe ciphertexts. In: *USENIX security symposium*, vol 2011, no 3
5. Porwal S, Mittal S (2017) Implementation of ciphertext policy-attribute based encryption (CP-ABE) for fine grained access control of university data. In: *2017 tenth international conference on contemporary computing (IC3)*. IEEE, pp 1–7

6. Bethencourt J, Sahai A, Waters B (2007) Ciphertext-policy attribute-based encryption. In: 2007 IEEE symposium on security and privacy (SP'07). IEEE, pp 321–334
7. Hasegawa K, Kanayama N, Nishide T, Okamoto E (2016) Software library for ciphertext/key-policy functional encryption with simple usability. *J Inf Process* 24(5):764–771

Restore Fingerprints Using Pix2Pix



Ji-Hwan Moon, Jin-Ho Park, and Gye-Young Kim

Abstract Previously studied fingerprint readers usually used Minutiae feature. Minutiae uses both directional maps and skeleton images because of its high FMR (False Match Rate). But unlike security attacks on Minutiae, research on directional maps and skeletal image attacks is not going well. In this paper, fingerprint images are generated using the new Pix2Pix model and analyzed representation attack vulnerabilities for the images. When the restored fingerprint by the model was recognized to the fingerprint recognizer, it showed a high recognition success rate and demonstrated the vulnerability for representation attacks on fingerprint readers that also use skeletal images.

Keywords Fingerprint · Fingerprint reader · Representation attack · Pix2Pix · Bio-feature

1 Introduction

Fingerprint recognition technology, which is used as one of the biometric recognition technologies, is the most commonly used biometric recognition technology due to the high universality, permanence, and acquisition of fingerprints [1]. Fingerprint recognition devices store fingerprints in a template format. Typical templates used in fingerprint readers include orientation maps, skeleton images, and minutiae, which show the ridges of fingerprints [2], as shown in Fig. 1. Unlike other features, Minutiae is often used in current fingerprint readers because the information is so sparse that

J.-H. Moon · J.-H. Park · G.-Y. Kim (✉)
Soongsil University, 378, Sangdo-ro, Dongjak-gu, Seoul, Republic of Korea
e-mail: gykim@ssu.ac.kr

J.-H. Moon
e-mail: gkrrydn_ji@naver.com

J.-H. Park
e-mail: j.park@ssu.ac.kr

Fig. 1 Fingerprint features used as templates



it cannot be used as a representation attack, making it safe enough to be adopted as an ISO/IEC standard biometric technology template.

With the recent development of computer performance, research is underway to restore fingerprint images using neural networks. In particular, the GAN (Generative Adversarial Network) [3] is a typical neural network-based generation model that consists of a generator that generates data and a discriminator that obscures the authenticity of the data. Bontrager [4] proposed a method using Wasserstein GAN (WGAN) [5], an extension of the GAN, to create a master fingerprint that could invalidate the fingerprint recognition device. Lee [6] proposed a method of restoring fingerprints from type analysis features through the CGAN (Conditional GAN) [7] and showed vulnerability to representation attack on fingerprint recognition devices that use fingerprint type features as templates.

Fingerprints restored by the template security vulnerability study of existing fingerprint recognizer are artificial and have unnatural form, so the success rate of representation attack is not high. Therefore, the purpose of this paper is to use the Pix2Pix model to restore a fingerprint similar to the actual fingerprint from the skeleton image of the fingerprint to deceive the discriminating system of the fingerprint recognizer. In addition, we want to study the security vulnerabilities of fingerprint readers by showing a higher success rate of representation attacks than previous studies.

2 Skeleton to Fingerprint Using Pix2Pix

The proposed method is largely divided into four by fingerprint feature extraction, fingerprint restoration model learning, fingerprint restoration, and restoration performance evaluation, and the schematic diagram are shown in Fig. 2. In the fingerprint characterization extraction stage, the experiment is carried out under the assumption that the skeleton images recovered from Minutiae are perfect. Fingerprint recognition devices perform erosion and dilation operations on images to remove dust-like noise after scanning fingerprints. Then, binary thresholding like Otsu is applied to create black-and-white images, and then a skeleton image is created through thinning. In the process of training the fingerprint restoration model, bone images are trained using the Pix2Pix model. Pix2Pix is an image-to-image conversion model using a

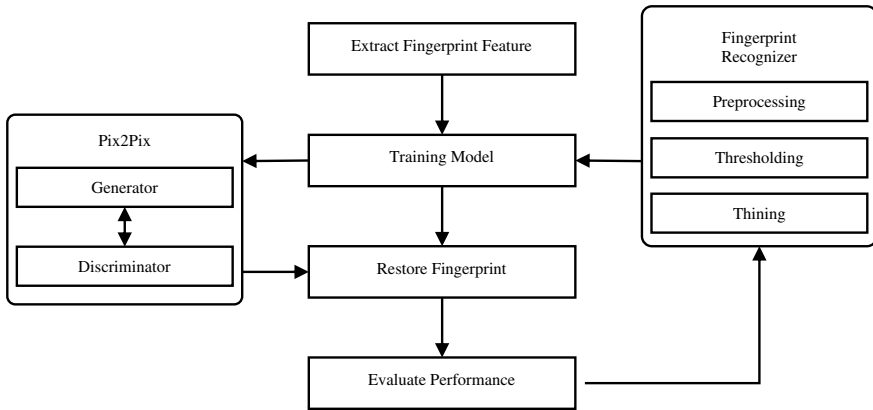


Fig. 2 Overview of fingerprint restores method

neural network that can be converted from input images to images of different colors or styles [8]. While normal GAN is difficult to predict output results due to output data from Hyperparameter, Pix2Pix has the advantage of determining the output for the input.

The neural network structure of the proposed model for restoring fingerprints from skeleton images is shown in Fig. 3. The model uses the Skip-Connection Encoder-Decoder used in U-Net [9], which internally enters 256×256 images and outputs 3 channel 256×256 images as shown in Fig. 3. A typical Encoder-Decoder loses many features in the data coding process, but Skip-Connection does not damage the features and moves on to the next tier. In addition, the neural network of the model uses 3×3 Convolution, Transposed Convolution Filter, and ReLU [10] as an activation function. The last result output layer uses Hyperbolic Tangent.

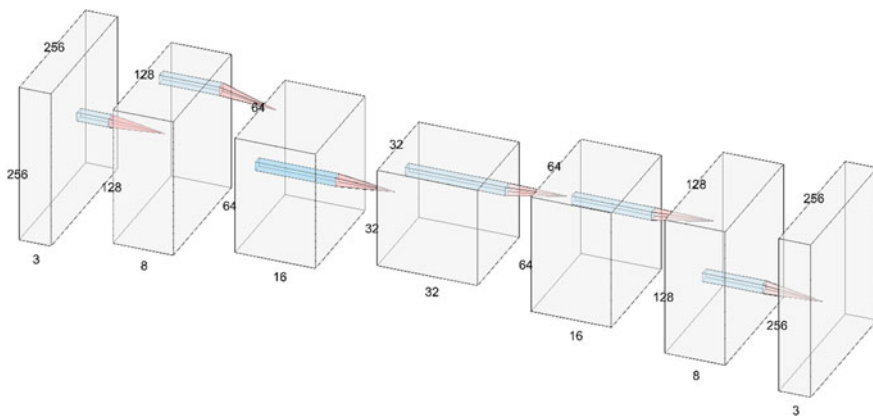


Fig. 3 Proposed neural network of Pix2Pix model

In the stage of fingerprint restoration and performance evaluation, skeleton images are entered in the trained Pix2Pix model to restore the fingerprint. As a stage of the restoration performance evaluation, the degree of similarity between restored and actual fingerprints is measured, and recognition accuracy is measured using a fingerprint recognizer. In the experiment, FID (Fréchet Inception Distance) [11] was used to compare similarities.

3 Experiment and Result

The experiment was conducted using Intel i9-7980XE CPU and two Titan-V GPUs. Fingerprint images used in the experiment were used by NIST Special Database 4, 3000 of 3,844 images were used for learning, 744 as validation sets and 100 as test sets.

The experiment studied the cGAN model used as a conditional generation model and the proposed Pix2Pix model which was carried out by changing the hyperparameter reconstruction variables and the reconstruction weights. Figure 4 is a fingerprint image restored according to the hyperparameter of the cGAN model and the Pix2Pix model.

For a well-learned GAN, it is known that the FID for one-channel images is 17–30 and the FID for three-channel images has a value of 103–192 [12]. Looking at Table 1, the FID of the cGAN model is 380.1668 showing very poor quality. Conversely, Pix2Pix shows very good quality between 105 and 171 except in some cases (L2). The test results show the best quality when using L1 along with reconstruction weight

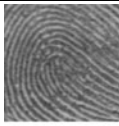
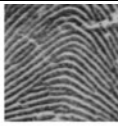
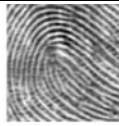
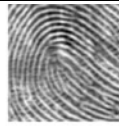

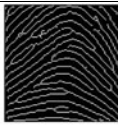
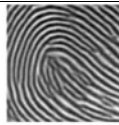
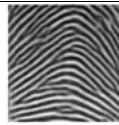
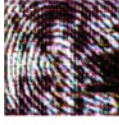

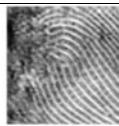
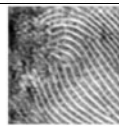

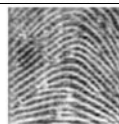
Input Image	Fing erprint			$\lambda = 50$	L1		
	Skeleton				L2		
cGAN				$\lambda = 100$	L1		
					L2		

Fig. 4 Proposed neural network of Pix2Pix model

Table 1 FID according to the hyperparameter of cGAN and proposed Pix2Pix model. L1 is norm1, L2 is norm2, both of norm is the reconstruction parameter of Pix2Pix. Lambda is the reconstruction weights of the Pix2Pix model

	L1 Loss	L2 Loss
cGAN	380.1668	
Proposed ($\lambda = 50$)	139.6825	402.1813
Proposed ($\lambda = 100$)	105.0424	118.5930

Table 2 The recognition rate of restored fingerprints according to FMR of fingerprint recognizer

	FMR 0.01 (%)	FMR 0.1 (%)	FMR 1 (%)
cGAN	0	0	1
Proposed ($\lambda = 50$) + L1	60	92	100
Proposed ($\lambda = 50$) + L2	69	94	100
Proposed ($\lambda = 100$) + L1	58	92	100
Proposed ($\lambda = 100$) + L2	62	86	100

100. Table 2 shows the success rate of the representation attack by recognizing restored fingerprints on the actual fingerprint recognizer according to the FMR. If the fingerprint is restored to the cGAN model, it shows that it has almost failed to attack the expression. For the proposed Pix2Pix model, it appears that there is no correlation between the success rate of the representation attack and the variables of reconstruction.

4 Conclusion

In this paper, the neural network of the Pix2Pix model is proposed to restore from the fingerprint skeleton image to the actual fingerprint, and to the fingerprint with a high success rate of representation attack. When fingerprint skeleton images were restored to the fingerprint from the proposed Pix2Pix model, they were sufficiently similar to the actual fingerprint through FID, which measures similarities with actual data. And use a fingerprint reader to show the current security vulnerability to fingerprint readers by showing a success rate of representation attacks that exceed 90% when FMR is 0.1%. Minutiae, a template considered safe, also requires a study to extract cross-certifiable features, not only those that are vulnerable to security, as the proposed method could restore Minutiae to a fingerprint from a skeleton image of the restored fingerprint.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01419) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. German RL, Barber KS (2018) Current biometric adoption and trends, A UT CID Report
2. Daugman JG (2004) How iris recognition works. *IEEE Trans Circuits Syst Video Technol* 14(1):21–30
3. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *NIPS*
4. Bontrager P, Togelius J, Memon N (2017) DeepMasterPrint: generating fingerprints for presentation attacks. <https://arxiv.org/abs/1705.07386v1>
5. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein gan. [arXiv:1701.07875](https://arxiv.org/abs/1701.07875)
6. Lee S, Jang SW, Kim D, Hahn H, Kim GY (2019) Synthesizing fingerprint from pattern type analysis features using cGAN, 19 April 2019
7. Mirza M, Simon O (2014) Conditional generative adversarial nets. <https://arxiv.org/abs/1411.1784>
8. Isola P, Zhu JY, Zhou T, Efros AA (2016) Image-to-image translation with conditional adversarial networks, <https://arxiv.org/abs/1611.07004>
9. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. <https://arxiv.org/abs/1505.04597>
10. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: *ICML'10 proceedings of the 27th international conference on international conference on machine learning*, pp 807–814
11. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter, S (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: *NIPS*, pp 6629–6640
12. Lee YJ, Seok KH A study on the performance of generative adversarial networks. *J Korean Data Inf Sci Soc* 1,155–1,167. <https://doi.org/10.7465/jkdi.2018.29.5.1155,2018.09>

Web Site Usage History Management System Using Blockchain



Cheolmin Yeom, Seonghwa Yeon, Sunghyun Yu, and Yoojae Won

Abstract Various types of data are being collected given the recent increase in the kinds of services provided by companies. Among them, data collectors of web services collect MyData indiscriminately and benefit from them. Moreover, data providers are unaware of how their data are collected and used. Data collectors of web services consume web resources by generating a large amount of web traffic in the process of data collection. This traffic can cause damage such as service interruption. In this study, we propose a web site search engine that builds a system to control user information using blockchain, and builds on the recorded information. This system allows data providers to manage MyData more transparently, and search engines using blockchain can contribute to creating a better web ecosystem.

Keywords MyData · Data sovereignty · Blockchain · EOS · Smart contract · Proxy server · Search engine

1 Introduction

As services using big data and artificial intelligence are developed and commercialized, the importance of data is increasing. In particular, the value of the user data has become more important as web services can generate profits through customized services or advertisements. These data are defined as MyData and can include data on

C. Yeom · S. Yeon · S. Yu · Y. Won (✉)

Department of Computer Science Engineering, Chungnam National University, Daejeon, South Korea

e-mail: yjwon@cnu.ac.kr

C. Yeom

e-mail: cjfals18@cnu.ac.kr

S. Yeon

e-mail: yeonseonghwa@cnu.ac.kr

S. Yu

e-mail: yoursaint@cnu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_69

an individual's purchases, communication, financial information, and online services used. The goal in this case is to ensure that the data are held and controlled by the individual [1, 2]. Such data should not be collected indiscriminately, and should recognize individual ownership [3]. However, data collectors of web services are typically unconcerned about privacy breaches caused by the use of personal data, and serious privacy risks can arise in cases of data over-collection [4, 5].

In addition, among web services, search engines generate a considerable amount of web traffic given their use of automated programs such as bots and web crawlers to collect data while providing search services to users [6]. As of 2018, web bots accounted for 42.2% of all website traffic [7]. This traffic causes much damage. For low specification servers with limited resources specifically, unnecessary resource consumption causes damage in the form of service interruption [8]. In order to solve this problem, it is important to study how the generation of unnecessary traffic may be avoided.

To this end, this study used blockchain to guarantee data sovereignty and solve traffic problems inherent in search engine use. Smart contracts allow users to control and collect MyData and increase efficiency by automating the processing of payments between users [9, 10]. We solved the problem of increased traffic by allowing the search engine to use information from the web blockchain network, which is based on collected information. In other words, it is possible to build a better web ecosystem by building a website search engine with the stored information.

2 Related Works

2.1 Blockchain

Blockchain is a public ledger of a distributed record database or events executed and shared by participants. Each transaction is verified by the participants, and the verified transaction cannot be tampered with. Blockchain is made of blocks that are linked together. A block consists of two parts: a block header and a transaction. The block header consists of six pieces of information: the hash value of the previous header, the version, and the time. If the block contains basic information, the transaction contains important information [9]. Smart Contract is a blockchain technology in the form of a computer program that can automatically execute the contract terms. If the preconfigured conditions are met, the smart contract is executed and the parties involved in the contract can proceed transparently [11, 12].

2.2 EOSIO

EOS is a third-generation cryptocurrency that uses Delegated Proof of Stake (DPoS). It is a decentralized platform based on blockchain technology. EOS is designed to support decentralized applications, and it is possible to develop smart contracts simultaneously without the burden of fees and about 200 times faster than when using second-generation cryptocurrency. In terms of usability, the complex address is changed into an easy-to-understand account, and in the case of Decentralized applications (DApp), it offers a faster processing speed than those of the other blockchains [13].

2.3 Blockchain Service

Blockchain and smart contracts can be applied to various services. For example, medical practitioners require an electronic medical record management system that provides comprehensive data on patients and shares medical information using the blockchain [14]. In addition, supply chain management systems in industries ensure the transparency of each transaction by applying smart contracts. The transactions are automatically stored on the blockchain network and can help the company save on administrative costs [15].

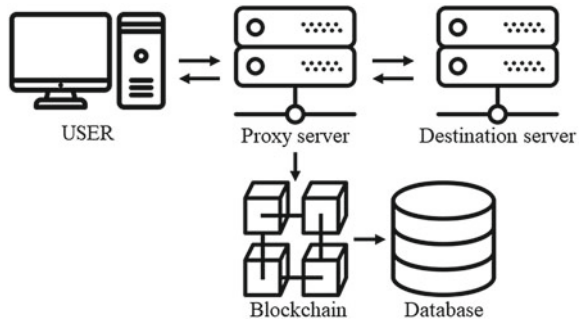
3 Web Site Usage History Management System

In this study, we propose a search system that manages MyData using blockchain and proxy servers, and resolves the problem of excessive traffic inherent to search engines.

3.1 Server Operation Process

As show in Fig. 1, a proxy server is used to store the user's web records in the block through a smart contract, and the search engine parses the web records stored in the blockchain network as search results. Rewards are forwarded to the personal information providers used as search results.

Fig. 1 Server operation process



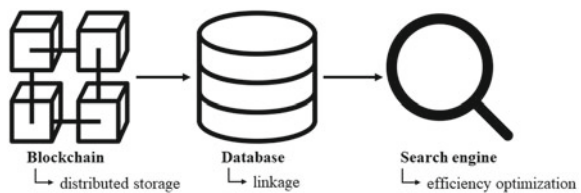
3.2 Search Engine Mechanism

The search engine Mechanism is shown in Fig. 2. The user can use the search function by selecting the desired node among several nodes of the blockchain network. Nodes in the blockchain network use smart contracts to write information to the blockchain, and block data are written to the database for efficient use. If the user has a token, he/she can vote for the node he/she wants depending on the token’s influence, and the node is then promoted to a block producer. The information stored in the database is sent to the search engine for use on the search engine site. If other providers’ personal information is used for personalized services, compensation will be delivered to the providers’ wallets through smart contracts.

4 Experiment

Users can create their own wallet account, which contains name and coin information. When creating a wallet and using a search engine, the search engine collects the user’s information, reflects it in the search results, and rewards the user. The search engine provides a search function and collects information about the user’s searches, his/her wallet name, search keyword, and site address, and stores these items in the block. In addition, users can mark pages they find useful as “Good” and use them to determine their usefulness.

Fig. 2 User information storage and use process



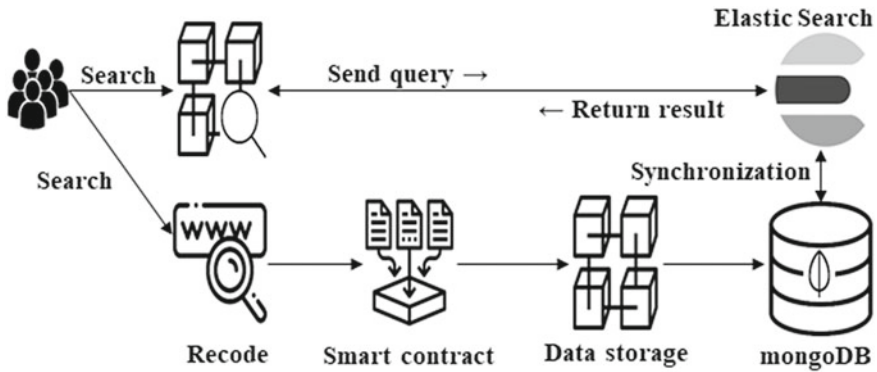


Fig. 3 Operation process

Table 1 Collected Information

Wallet name	Search keyword	Example URL	Number “Good”
A, C, D	Eos	eos.io	3
A, C	Eos	coinbase	2
A	Eos	eoshashnet	1
B	Eos	eoswiki	0

4.1 Scenario

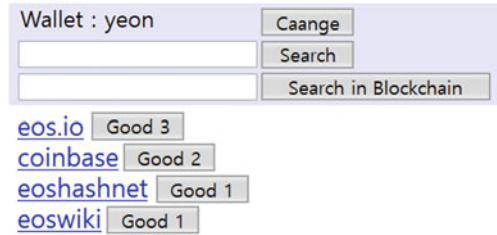
As shown in Fig. 3, when a user visits a desired site, the wallet name, search keyword, site address, and information about marked pages are collected and stored in the blockchain through Smart Contract. When a transaction occurs, each item of data is synchronized to the database. The user can search for the desired information on the blockchain. The search engine completes the search and returns the recommendations by order of relevance.

For the experiment, we searched for the keyword “Eos” using accounts A, B, C, and D, as shown in Table 1, and then hit “Good” on each page that was found.

4.2 Result

Searching for “Eos” within the blockchain with the new account provided the following results on the web in descending order of relevance with regard to the most “Good” ratings provided to the site: eos.io, coinbase, eoshashnet, and eoswiki. Giving “Good” to informative posts on the results page will reward the account that provides the document (Fig. 4).

Fig. 4 Searched data in the blockchain



Unlike the existing service, the proposed search engine can generate profits by employing users' own activity information, and the information storage management system based on blockchain ensures the safety and reliability of the information.

5 Conclusion

This paper proposes a system that records and utilizes users' MyData with blockchain. It describes the manner in which information may be managed, reasonable payments be made for the use of such information, unnecessary web traffic can be minimized for automated programs. Use of the blockchain allows MyData to be managed more transparently; its contents can be directly checked by users and the data will be sovereign.

The website search engine presented as a blockchain utilization method is built on the user's web usage history. It will contribute to creating a better web ecosystem by resolving the problem of indiscriminate traffic as it relies on automation programs for collecting existing web records. This aspect reduces the consumption of resources for service maintenance and provides more reliable service as it is constructed with user information.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00304) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

"This research was supported by the MIST (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & communications Technology Promotion)" (2015-0-00,930).

References

1. Poikola A, Kuikkaniemi K, Honko H (2015) Mydata—a nordic model for human-centered personal data management and processing. Finnish ministry of transport and communications
2. Häkkinen J, Alhonsuo M, Virtanen L, Rantakari J, Colley A, Koivumäki T (2016) MyData approach for personal health—a service design case for young athletes. In: 49th hawaii international conference on system sciences (HICSS). IEEE, Koloa, HI, USA
3. Kang R, Dabbish L, Fruchter N, Kiesler S (2015) My data just goes everywhere: user mental models of the internet and implications for privacy and security. In: Eleventh symposium on usable privacy and security (SOUPS) 2015). Usenix, Ottawa, Canada
4. Kamleitner B, Mitchell V (2019) Your data is my data: a framework for addressing interdependent privacy infringements. *J Public Policy Mark* 38:433–450
5. Dai W, Qiu M, Qiu L, Chen L, Wu A (2017) Who moved my data? privacy protection in smartphones. *IEEE Commun Mag* 55:20–25
6. Zineddine M (2016) Search engines crawling process optimization: a webserver approach. *Internet Res* 26:311–331
7. Distil networks: the 2018 bad bot report. <https://resources.distilnetworks.com/white-paper-reports/2018-bad-bot-report>
8. Owezarski P (2005) On the impact of DoS attacks on internet traffic characteristics and QoS. In: Proceedings of the 14th international conference on computer communications and networks, 2005 (ICCCN 2005). IEEE, San Diego, CA, USA
9. Nakamoto S Bitcoin: a peer-to-peer electronic cash system. www.bitcoin.org
10. Idelberger F, Governatori G, Riveret R, Sartor G (2016) Evaluation of logic-based smart contracts for blockchain systems. In: International symposium on rules and rule markup languages for the semantic web. Springer, Cham, Stony Brook, NY, USA
11. Crosby M, Nachiappan P, Verma S, Kalyanaraman V (2016) Blockchain technology: beyond bitcoin. *Appl Innov Rev* 6, 6–19
12. Christidis K, Devetsikiotis M (2016) Blockchains and smart contracts for the internet of things. *IEEE Access* 4:2292–2303
13. EOSIO developer portal. <https://developers.eos.io>
14. Azaria A, Ekblaw A, Vieira T, Lippman A (2016) MedRec: using blockchain for medical data access and permission management. In: 2nd International conference on open and big data (OBD). IEEE, Vienna, Austria
15. Yoo M, Won Y (2018) A study on the transparent price tracing system in supply chain management based on blockchain. *Sustainability* 10:4037

Generation of Fake Iris Images Using CycleGAN



Jae-gab Choi, Jin-Ho Park, and Gye-Young Kim

Abstract With the development of biometric recognition technology, identification of users through biometric information such as iris, fingerprint, and palm print is being applied to many areas. In the case of iris, various methods of recognition and methods of detection of fake iris have been studied at the iris recognition stage. However, fake iris detection research has been conducted by using the printed output or the artificial iris due to the absence of fake iris data. In this paper, fake iris images are generated for the research of the detection of fake iris using CycleGAN. The CycleGAN model has learned to reduce constraints against existing generation models and to avoid bias in probability distributions using bidirectional LossFunction. In the experiment, CASIA Iris Image Database ver 4.0 was used and the data was obtained for the detection of fake iris by creating fake iris.

Keywords GAN · CycleGAN · Iris · Fake image · Fake iris

1 Introduction

With the development of biometric technology, biometric information such as palm prints, fingerprints, and irises are used to identify users in various fields such as security, finance, and immigration. Iris discriminates against individuals compared to other living features, and is gaining popularity along with fingerprints. Identity identification of users using iris can be divided into iris image acquisition and iris recognition steps. In the iris image acquisition stage, images are acquired and stored through the camera. In the iris recognition phase, the characteristics of the iris area are

J. Choi · J.-H. Park · G.-Y. Kim (✉)
Sangdo-ro, Dongjak-gu, 378 Seoul, Republic of Korea
e-mail: gykim11@ssu.ac.kr

J. Choi
e-mail: kor_03@naver.com

J.-H. Park
e-mail: j.park@ssu.ac.kr

extracted by dividing the pupil and iris areas, and the characteristics are compared. During the iris recognition phase, numerous approaches have been studied to identify users using such methods as phase-based iris recognition [3] or Dogman's iris recognition [2]. In addition, SVMs [4], Purkinje phenomenon [5], iris encryption solution, and methods for identifying fake iris using fake iris detection [6] were studied. However, the fake iris detection study has a problem of using fake iris or artificial iris outputted from paper because there is no fake iris sample.

One of the unsupervised learning, the Generative Adversarial Network (GAN), was studied and used to enable the creation of fake images. Research has been done on the creation of fake iris images using GAN [7, 10–13]. In this paper, we propose a method of generating a fake iris image using CycleGAN, which is much less constrained by using the unpaired training set and prevents the noise deflection of the generator and discriminator in the existing GAN model.

In this paper, Sect. 2 describes the structure and experimental method of CycleGAN. In Sect. 3, the experimental environment setting and the fake iris generation and results through CycleGAN introduced in Sect. 2 are presented. Finally, data for iris discriminator learning is generated.

2 Original Iris Image-to-Fake Iris Image Translation

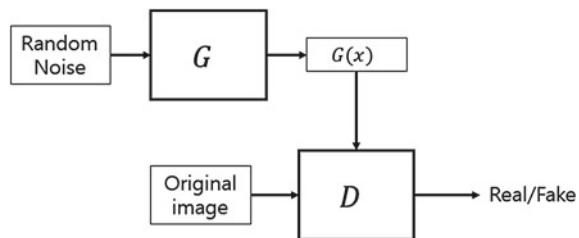
2.1 Generative Adversarial Network

GAN is one of neural network based generation models [9]. As shown in Fig. 1, the Generator (G : Generator) and the Discriminator (D : Discriminator) repeatedly learn and derive the optimal solution.

Generator G as shown in Fig. 1 enters the random noise z generated through the probability distribution to produce a fake image. Discriminator D determines whether the entered data is the generated image or the original image. The objective function of the GAN is shown in formula (1).

$$\min_G \max_D V(D, G) = E_{x \sim P_d(x)}[\log D(x)] + E_{z \sim P_d(z)}[\log(1 - D(G(z)))] \quad (1)$$

Fig. 1 Generative Adversarial Network Architecture



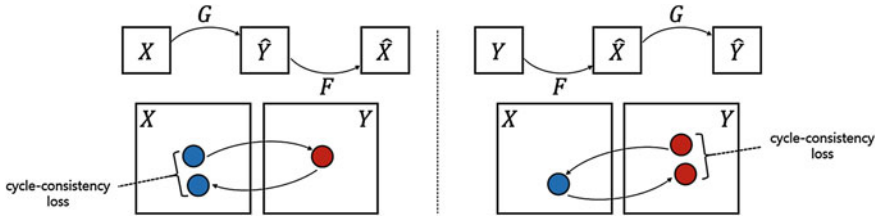


Fig. 2 CycleGAN circulation structure

These neural network-based production models allow the creation of fake iris images, and the absence of fake iris samples from previous studies can be solved by utilizing the fake iris images produced.

2.2 CycleGAN

CycleGAN is one of the models of the General Adaptive Advertising Network. Figure 2 shows the model structure of the CycleGAN and uses the unpaired training set. The use of unpaired training sets has eliminated the restriction on the use of paired training sets in existing. CycleGAN also has the advantage of using U-Net to retain more detail of the generated images.

$$\begin{aligned}
 Loss_{x \rightarrow y} &= E_y[\log(D_y(y))] + E_x[\log(1 - D_y(x))] + E_x[\|F(G(x)) - x\|_1] \\
 Loss_{y \rightarrow x} &= E_x[\log(D_x(x))] + E_y[\log(1 - D_x(y))] + E_y[\|F(G(y)) - y\|_1]
 \end{aligned} \tag{2}$$

$$Loss_{cycleGAN} = Loss_{x \rightarrow y} + Loss_{y \rightarrow x}$$

Formula (2) is the Loss Function of CycleGAN. The image quality was improved by creating images using bidirectional Loss Function. This prevents the noise of probability distribution from being biased due to repeated learning of Generator and Discriminator. Figure 3 below shows the bias of noise.

3 Create Fake Iris Image

In this chapter, fake iris images are created to overcome the absence of samples in the fake iris detection study. CycleGAN as described in Sect. 2 was used to generate fake iris images. CASIA Irisimage Database ver 4.0 was used to generate fake iris images. In addition, arbitrary iris-shaped edge images were used as input images. Since imaging is well produced at $\lambda \geq 10$, the experiment was conducted using the specified λ of 10. Figure 4. shows a fake iris image produced through an experiment.

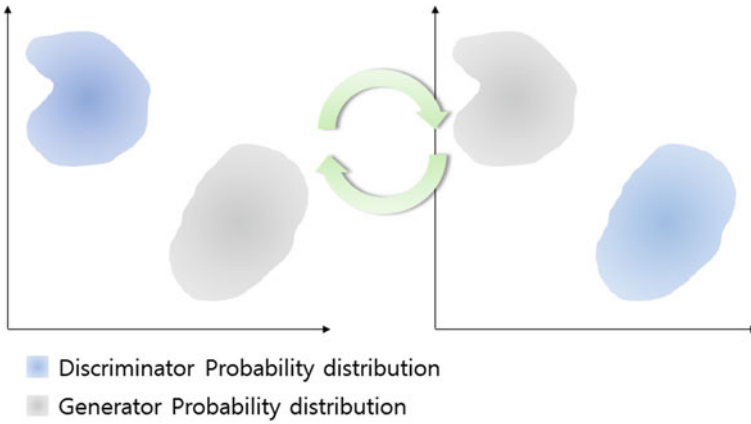


Fig. 3 Probability Distribution Map for Generator and Discriminator

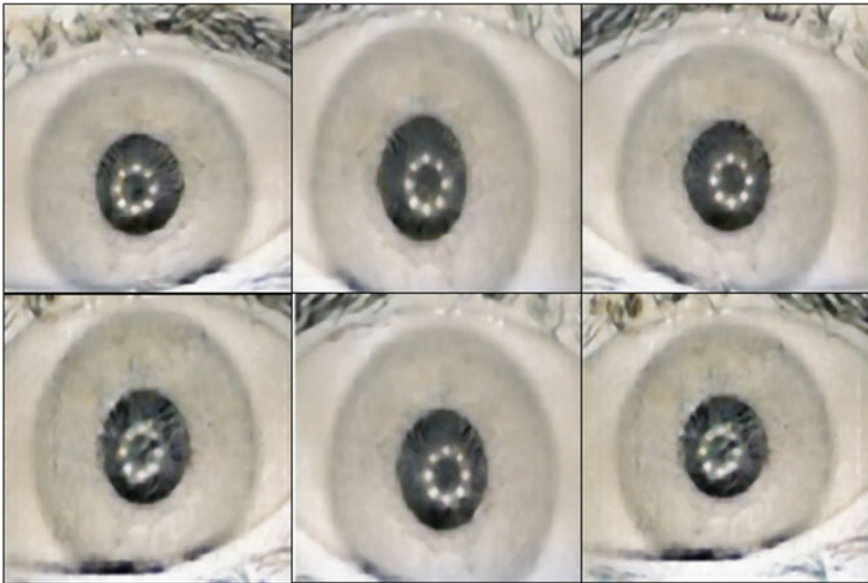


Fig. 4 Generated fake iris image

1 The time per EPOCH was approximately 350 s and it took about 2 days for the entire learning process.

In this paper, the advantage is that fake iris images were created using CycleGAN to enable the creation of fake iris by a small number of domains. However, the fake iris has the disadvantage of having visually discernable noise.

4 Conclusion

In this paper, the method of producing fake iris images was proposed using CycleGAN, one of the models of the Generative Adversarial Network. Through experiments, we could generate data for learning iris discriminator. Using CycleGAN, noise was not skewed to one side in the probability distribution of the neural network, and the restriction on the use of pairing training sets of existing generation models was reduced to create fake iris image. Also, through the creation of fake iris images, we took many samples from the research to detect fake iris and made it possible to experiment. In order to produce high quality fake images in the future, it is necessary to continuously check quality by adjusting the parameters of the back and the neural network structure.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01419) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Jain A, Flynn P, Ross A (Eds) Handbook of biometrics. Springer
2. Daugman J (2006) Probing the uniqueness and randomness of iriscodes: results from 200 billion iris pair comparisons. Proc IEEE 94 (11):1927–1935
3. Miyazawa K, Ito K, Aoki T, Kobayashi K, Nakajima H (2006) A phase-based iris recognition algorithm. Adv Biom 3832:356–365
4. He X, Lu Y, Shi P (2009) A new fake iris detection method ICB 2009: advances in biometrics, pp 1132–1139
5. Lee EC, Park KR, Kim J (2006) Fake iris detection by using purkinje image ICB 2006: advances in biometrics, pp 397–403
6. Sinha VK, Gupta AK (2018) Manish mahajan: detecting fake iris in iris biometric system. Digit Investig 25:97–104
7. Minaee S, Abdolrashidi A (2018) Iris-GAN: learning to generate realistic iris images using convolutional GAN. [arXiv:1812.04822](https://arxiv.org/abs/1812.04822)
8. Zhu JY, Park T, Isola P, Efros AA (2017) Un paired image-to-image translation using cycle-consistent adversarial networks. CoRR, abs/1703.10593
9. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. <https://arxiv.org/abs/1406.2661>
10. Lee S, Jang SW, Kim D, Hahn H, Kim GY (2019) Synthesizing fingerprint from pattern type analysis features using cGAN. 19 April 2019
11. Xiao T, Hong J, Ma J (2018) DNA-GAN: learning disentangled representations from multi-attribute images. In: Proceedings of the international conference learning representing (ICLR) workshops
12. Chang H, Lu J, Yu F, Finkelstein A (2018) PairedCycleGAN: asymmetric style transfer for applying and removing makeup. In: Proceedings of the IEEE computer conference on computer vision and pattern recognition (CVPR), pp 40–48
13. Lu Y, Tai Y-W, Tang C-K (2018) Attribute-guided face generation using conditional CycleGAN. In: Proceedings of the European conference on computer vision (ECCV), pp 282–297

Robust 3D Reconstruction Through Noise Reduction of Ultra-Fast Images



Nu-lee Song, Jin-Ho Park, and Gye-Young Kim

Abstract 3D reconstruction from multiple view images has been studied extensively in computer vision tasks. In order to increase the accuracy of the 3D reconstruction, it is important to secure the number of image frames and to find feature points and match accurate feature points by minimizing the influence of noise from each image. When we acquired images from high-speed camera, it is possible to analyze phenomena and object movements that are difficult to see with the naked eye. However, when using a high-speed camera, problems such as increased data amount, light amount, focus, and noise occur due to an increase in resolution and shutter speed. In this paper, we propose a preprocessing method for feature point tracking and matching for robust 3D reconstruction in high-speed images. The experimental results confirm the validity compared with 3D reconstruction output from the original image and preprocessed image.

Keywords 3D reconstruction · High-speed image · Image procession

1 Introduction

A study of 3D reconstruction from 2D images has been conducted in the field of computer vision, the actual estimation result is unsatisfactory because of distortion caused by noise included in the measured data [1]. Noise is generated from various causes such as signal interference and deterioration in the photographing environment, sensor, and transmission process during image acquisition [2].

N. Song · J.-H. Park · G.-Y. Kim (✉)
378, Sangdo-ro, Dongjak-gu, Seoul, Republic of Korea
e-mail: gykim11@ssu.ac.kr

N. Song
e-mail: nuri@soongsil.ac.kr

J.-H. Park
e-mail: j.park@ssu.ac.kr

3D reconstruction techniques include structured light techniques, three-dimensional laser scanning, and methods to find and reconstruct feature points in an image without artificially projecting energy onto the object. The structured light technique uses a projector to project a predefined pattern onto an object to calculate the amount of change and restore the three-dimensional shape [3]. The laser scanning method performs modeling by projecting a single line or multiple lines of lasers onto an object [4]. Structure from Motion (SFM) is a typical method of 3D reconstruction using only images, and SFM finds key points in multiple images and matches the same key points. Then, to increase the density of the points, image-based modeling is completed by increasing the number of points using the Multi-View Stereo (MVS) algorithm, generating meshes and mapping textures to the model with the points [5].

When acquiring an image from a high-speed camera, blurring does not occur compared to a general image, a recognition ability of a moving object is high, and there is high visibility of the cause and process of the phenomenon [6]. When 3D reconstruction of a moving object using a high-speed camera is performed, the motion information of the object over time may be gradually updated using the information of previously acquired data.

High-speed shooting is affected by shooting speed, camera resolution, lighting, etc., and determines how sharp and bright the image will be taken according to the shooting speed. In the case of ultra high-speed photography, the faster the shutter speed, the lower the exposure value to the sensor, and thus the insufficient amount of light [7]. In this case, the image contains a lot of noise, which makes it difficult to obtain feature points for 3D reconstruction. Therefore, in this paper, we propose a noise reduction method of high-speed images for robust 3D reconstruction.

The proposed method acquires images with a high-speed camera. In the acquired image, we can confirm that the noise due to the fast shutter speed was distributed throughout the image. To solve this problem, we propose the method using the k-means algorithm. k-means clustering is used to remove the noise of the whole image. And to remove the outlier that was not removed, we adapt the optical flow to identify the motion of the feature point and calibration the value of the feature.

2 Image Calibration Using k-Means Clustering and Optical Flow

Figure 1 presents the flow chart of the proposed method.

In the proposed method, images are acquired from a fixed high-speed camera. The lighting should be as bright as possible and the distance between the cameras should be as narrow as possible to minimize noise and extract feature points for 3D reconstruction. However, basically, the images obtained from the high-speed images contain a lot of noise, and the result is shown in Fig. 2. Figure 2 shows the noise output from the difference between the random time t and $t + 1$ in the obtained high-speed image.

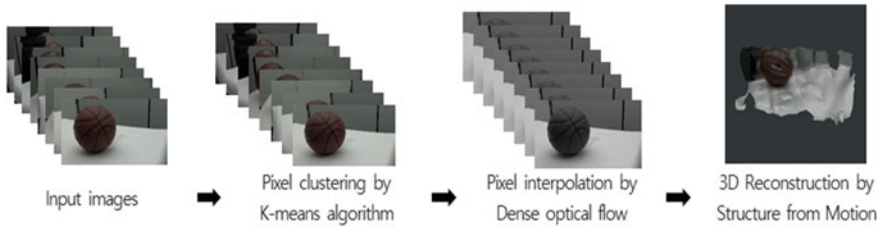


Fig. 1 Flow chart of the proposed method

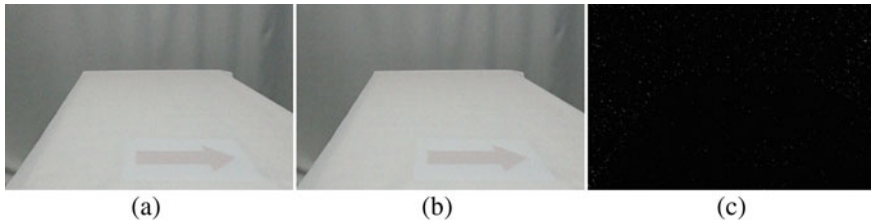


Fig. 2 a Time t image, b time t + 1 image, c difference image between (a) and (b)

As shown in (c) of Fig. 2, in a high-speed image, a lot of noises are distributed throughout the image, and the noise makes it difficult to accurately match the feature points between the images. For 3D reconstruction, a method of removing noise without damaging the feature points of the image should be used. In this paper, we propose the using k-means algorithm to remove the noise [8]. We can make it not recognize the noise that distributed throughout the image as the feature point without damaging feature points. If the number of clusters is small, all the features may disappear, so the number of clusters is increased so that the noise can be eliminated without disappearing as much as possible. At this time, the number of cluster k is selected empirically by roughly considering the color distributed in the image.

Even when k-means clustering is applied, the optical flow can be incorrectly estimated. In addition, in order to remove the remaining noise of images that have not been removed, we interpolated the pixel value using the flow of pixels in the image sequence. In this paper, we used dense optical flow to estimate the motion between images. Using dense optical flow estimates motion for every pixel, which makes motion estimation and pixel interpolation more accurate than Lucas-kanade optical flow of measuring motion for feature points [9].

When the motion of a pixel is estimated by using a dense optical flow in a group of five frames based on a specific frame in the image sequence, it is found that a pixel value in a specific frame is different from that of another frame, which is likely to be the outlier. Therefore, as a method for correcting this, the interpolation is performed by averaging the values of the pixels estimated in the front and rear time, as shown in Eqs. 1 and 2.

$$\left| \frac{p_{t-2} + p_{t-1} + p_{t+1} + p_{t+2}}{4} \right| > threshold \quad (1)$$

$$p_t = \frac{p_{t-1} + p_{t+1}}{2} \quad (2)$$

Finally, 3D reconstruction is performed using the structure from motion (SFM) method using the preprocessed image in the above process.

3 Experimental Results

In order to implement the proposed method and evaluate the performance, Intel Core I7-7700 (3.6 GHz) and input images are acquired by Sony DSC-RX100M6 camera. The acquired images are capable of shooting up to 960 fps per second at a resolution of 1920×1080 . In this experiment, 8 fixed high-speed cameras were used, and Fig. 3 is a multi-view image obtained at an arbitrary time point by the high-speed camera.

Figure 4 shows the results of three-dimensional reconstruction before and after preprocessing in ultrafast images. Figure 4a shows the existing input image and (b) shows the result of applying to preprocess.

When 3D reconstruction is performed using the original images obtained by the ultra high-speed camera, many unreconstructed regions appear in the 3D reconstruction image. However, when the preprocessing and 3D reconstruction are performed using the proposed method, the noise reduction of the image is reduced, and the feature matching rate is improved by about 10% than when the original image is reconstructed.



Fig. 3 High-speed images at random time T

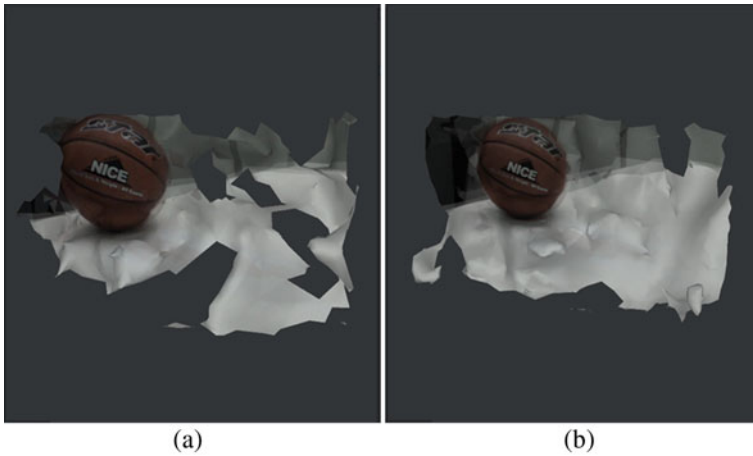


Fig. 4 **a** 3D restored image before preprocessing, **b** 3D restored image after preprocessing

4 Conclusion

In this paper, we propose a 3D reconstruction method by minimizing the noise of images obtained from high-speed cameras. Since high-speed images are affected by noise and light, it is difficult to extract robust feature points to lighting and noise and perform the 3D reconstruction. In order to solve this problem, in this paper, we are applying a k-means method and dense optical flow to image to remove the noise. 3D reconstruction results are smoother and more precisely restored than when using the original image. The proposed method improves three-dimensional reconstruction accuracy by increasing the number of feature point matchings, but it takes a long time. We plan to supplement the research and experiment about moving objects.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01419) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Koh S-S (2016) Uncertainty analysis of observation matrix for 3D reconstruction. *J Korea Inst Inf Commun Eng* 20(3):527–535
2. Baek S, Jeong S, Choi J-S, Lee S (2015) Effective noise reduction using STFT-based content analysis. *J Inst Electron Inf Eng* 52(4):145–155
3. Pueo B (2016) High speed cameras for motion analysis in sports science. *J Hum Sport Exerc* 11:53–73
4. Pages J, Salvi J, Garcia R, Matabosch C (2003) Overview of coded light projection techniques for automatic 3D profiling. In: *IEEE international conference on robotics and automation*

5. Park H, Lee D (2018) 3D spatial data generation and data cross-utilization for monitoring geoparks: using unmanned aerial vehicle and virtual reality. *J Geol Soc Korea* 54(5):501–511
6. Fu Y, Liu Y (2018) 3D bubble reconstruction using multiple cameras and space carving method. *Meas Sci Technol* 29(7)
7. Koutsoudis A, Vidmar B, Loannakis G, Arnaoutoglou F, Pavlidis G (2014) Multi-image 3D reconstruction data evaluation. *J Cult Herit* 15, 73–79
8. He K, Wen F, Sun J (2013) K-means hashing: an affinity-preserving quantization method for learning binary compact codes. In: *Proceedings of IEEE conference on computer vision and pattern recognition*, pp 2938–2945
9. Walker J, Gupta A, Hebert M (2015) Dense optical flow prediction from a static image. In: *IEEE International Conference on Computer Vision (ICCV)*

Pedestrian Detection Using Regression-Based Feature Selection and Disparity Map



Chung-Hee Lee

Abstract In this paper, the pedestrian detection using a regression-based feature selection and a disparity map method is proposed for improving the processing speed. Using many features helps to improve detection performance, but slows down processing. Therefore, it is important to select and use features efficiently. Our proposed method consists of three stages, such as a disparity map-based detection stage, a segmentation stage using a transformed disparity map, and a recognition stage with regression-based feature analysis. Through experiments with the ETH database, we show that the proposed method improves detection performance and especially processing speed.

Keywords Detection · Regression · Disparity map

1 Introduction

Pedestrian detection using images is one of the important technologies required for various intelligent systems [1–4]. In particular, image sensors are applied to various fields because they provide various information as a single sensor unlike other sensors. Recently, with the development of various artificial intelligence algorithms, the performance of image recognition is improving [3, 4], and its importance is expected to increase further. However, since the image has a relatively large amount of data compared to other sensors, it takes a long time to process the image. Of course, hardware and software for processing image have been developed recently, but processing speed is still a big issue. In the future, if a lot of image recognition techniques using higher resolution images are used, the processing speed problem will be more highlighted. And one good way to improve the performance of obstacle detection is to use a disparity map [5–8]. The disparity map is an image of distance information expressed in the gray level of brightness. It is obtained by calculating the disparity value using a stereo matching algorithm in the left and right images. Because

C.-H. Lee (✉)

Daegu Gyeongbuk Institute of Science & Technology, Daegu, South Korea
e-mail: chlee@dgist.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_72

515

the disparity map provides three-dimensional information, it helps to improve detection performance by increasing the separation between obstacles and backgrounds. In addition, the disparity value can be utilized to estimate the distance from the obstacle. The most time-consuming part of image recognition is image classification. In general, in order to detect a specific obstacle in an image, image classification is performed on the entire image. If image classification is performed only within certain areas of the image, the overall image recognition speed will be significantly improved. The disparity map helps to find candidates for obstacle detection. Thus, the disparity map improves the processing speed by reducing the image area for image classification. And another way to improve processing speed in image classification is to use features efficiently. Using many features helps to improve detection performance, but slows down processing. Therefore, it is important to select and analyze features.

In this paper, the pedestrian detection using regression-based feature selection and disparity map method is proposed for improving the processing speed. Our proposed method consists of three stages, such as a disparity map-based detection stage, a segmentation stage using a transformed disparity map, and a recognition stage with regression-based feature analysis. In the first detection stage, all pedestrian candidates are detected using a v-disparity map and road feature information. The v-disparity map is obtained by histogramizing the disparity value of each row in the vertical direction of the disparity map [6, 7]. The v-disparity map represents obstacles on the road very well, which helps to detect obstacles easily. In the segmentation stage, the process of separating the detected obstacle areas more precisely is performed in order to solve the problem that the same obstacle is divided or the different obstacles are detected as one obstacle. The problem is caused by the disparity value error and the limitation of detection performance. In the final stage, the classification process is performed for each segmented area. When features are used in the classifier training step, not all features are used, only features that make a high contribution to pedestrian detection are used. When selecting features used in the classifier, a regression-based feature analysis is used. Through the regression method, we identify the contribution of specific feature element to classification.

2 Pedestrian Detection Using Regression-Based Feature Analysis and Disparity Map

2.1 Stereo Vision System Modeling

A disparity map is obtained by calculating the difference of correspondence points between left and right images captured by stereo camera. Because the disparity map offers three-dimensional information, the obstacle detection performance is improved and 3D obstacle tracking is also possible, which helps to improve overall detection performance. Figure 1 shows the modeling of a stereo vision system. Coord-

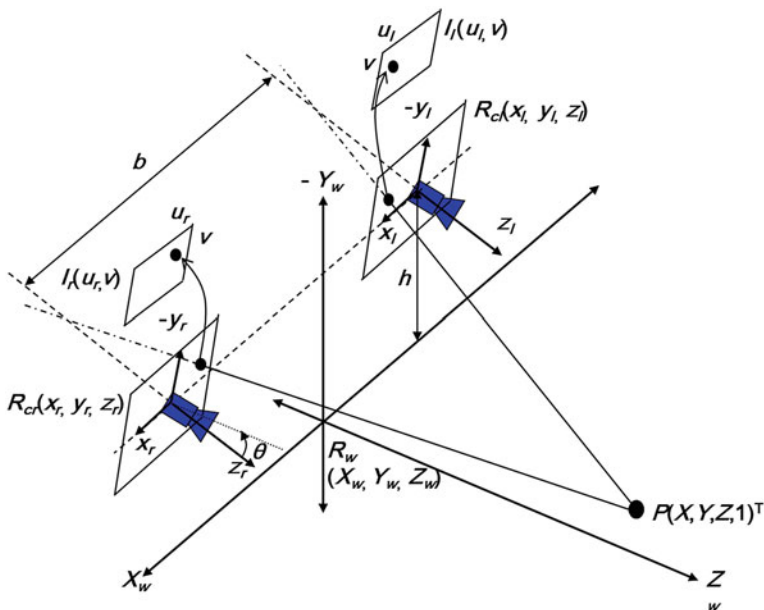


Fig. 1 The modeling of stereo vision system

dinates are composed of world coordinate system R_w , left and right camera coordinate systems R_{cl} , R_{cr} and image coordinate system (u, v) [6, 7]. We define a baseline, b , the distance between two cameras, and assume that each camera is located h from the ground on the world coordinate system. It is also assumed that both cameras are rotated by θ with respect to the X_w axis, which is the horizontal axis of world coordinates. Finally, if the center point of the image coordinate is defined as (u_0, v_0) , any point $P(X, Y, Z, 1)^T$ on the world coordinate is mapped to a point in the image coordinate through several steps. The mapping equations and disparity value, d are as follows.

$$u_l = u_0 + \alpha \frac{X + b/2}{(Y + h) \sin \theta + Z \cos \theta} , \tag{1}$$

$$u_r = u_0 + \alpha \frac{X - b/2}{(Y + h) \sin \theta + Z \cos \theta} , \tag{2}$$

$$v = v_0 + \alpha \frac{(Y + h) \cos \theta - Z \sin \theta}{(Y + h) \sin \theta + Z \cos \theta} , \tag{3}$$

$$d = u_l - u_r = \alpha b \frac{1}{(Y + h) \sin \theta + Z \cos \theta} . \tag{4}$$

where, α is focal length expressed by the number of pixels.

2.2 Obstacle Detection and Segmentation

The v-disparity map is generated from the disparity map [7], and the road feature information is extracted. Column detection is performed based on the extracted road feature information and various parameters [6]. Because the road feature information is used as a criterion for obstacles, it is important to extract the road feature information robustly. It is also necessary to consider various situations of obstacles and backgrounds to further improve the obstacle detection performance. All columns of the disparity map are compared with the road feature information. Since the disparity value of all obstacles on the road is larger than the value of the road feature information, if the row value in each column is larger than the value of the road feature information, the row is recognized as an obstacle. However, in consideration of the obstacle height and the disparity error, if there is a continuous row interval larger than the value of the road feature information, the interval is recognized as an obstacle. Although obstacle detection is performed using the disparity map, because of the disparity value error and large detection area, it is highly likely that a multiple obstacles are still included in a single obstacle area. Therefore, obstacle segmentation is performed to accurately detect as a single obstacle. Segmentation on the disparity map of the X - Z plane view has better performance and fast processing speed than segmentation on the X - Y plane view. Therefore, the X - Y plane view is transformed into the X - Z plane view. The mapping equations are as follows from Eqs. (1) to (3).

$$X = \{(Y + h) \sin \theta + Z \cos \theta\} \frac{u_l + u_r - 2u_0}{2\alpha}, \quad (5)$$

$$Y = 0, \quad (6)$$

$$Z = v_0 + \alpha \frac{(v_0 - v)b \sin \theta + \alpha b \cos \theta}{d}. \quad (7)$$

And then, grouping that combines neighboring rows is performed in the transformed disparity map. In consideration of the characteristic of the disparity value, the points with large disparity value merge with a large number of neighboring rows, and the points with small disparity value merge with a small number of neighbor rows. Histograms are generated for each group, and segmentation is performed. Finally, the histogram on the X - Z plane is inversely converted into a disparity map on the X - Y plane.

2.3 Regression-Based Feature Selection

Various kinds of features are used to improve image recognition performance. Using many types of features helps to improve recognition performance, but slows down

overall processing speed. Therefore, it is necessary to select and use features efficiently through analysis of the features. In other words, it is important to grasp the importance of each element and use features selectively without using all elements of the extracted features. The criterion for distinguishing between good and bad features is the ability to distinguish the difference between true and false learning images. For each feature used, the importance of each element can be determined by accumulating the true and false learning images. The classifier is trained using only selected features and tested using validation images. Upon verification of the classifier, four kinds of groups, such as true positive group, false negative group, true negative group, and false positive group, are generated. True positive and true negative groups are groups of correct classification results, but false negative and false positive groups are not. This represents the limit of the classifier consisting of the features selected in the previous step. Thus the better features need to be selected through analysis of these groups. Although false negative group must belong to true positive group, the two groups are separated due to a recognition error. Therefore, we assume the two groups as different groups and perform the feature analysis using regression on the used features. The high weighted elements of each feature are removed and new elements are selected. A new classifier is created using the new selected features and reverified using the verification images. Through this iterative process, the final classifier with optimal features is created.

3 Experiments

ETH database is utilized to verify the proposed algorithm [9]. Since ETH database provides stereo images, disparity maps can be generated using various stereo matching algorithms. We use a belief propagation algorithm, one of the global matching algorithms [10]. INRIA and our database are used for training and validation images. Color self similarity, histogram of oriented gradients and symmetry are used as features for experiments. First, in the feature cumulative analysis, the number of first selected features is 65% of the total number of features. This value generally affects learning speed and recognition performance. The most used index, the miss rate in false positive per image is used as a detection performance index. In particular, the proposed algorithm and the method using all the features are compared in terms of detection performance and processing speed. The detection performance of the proposed algorithm is 19%, which is 4% better than the method using all the features. When no hardware or software acceleration system is used, the processing speed is 47 s, roughly three times faster than all feature usage methods of 129 s.

4 Conclusions

In this paper, the pedestrian detection using regression-based feature selection and disparity map method was proposed for improving the processing speed. Our proposed method consisted of three stages, such as a disparity map-based detection stage, a segmentation stage using a transformed disparity map, and a recognition stage with regression-based feature analysis. In the first detection stage, all pedestrian candidates were detected using a v -disparity map and road feature information. In the segmentation stage, the process of separating the detected obstacle area more precisely was performed. In the final stage, the classification process was performed for each segmented area. When selecting features used in the classifier, a regression-based feature analysis was used. Through the regression method, we identified the contribution of specific feature element to classification. Through the experiments, The detection performance of our proposed algorithm was 4% better than the method using all the features and the processing speed was roughly three times faster than all feature usage methods. In the future, we will try to solve the mismatching problem between classifier and regression in feature selection.

Acknowledgements This work was supported by the DGIST R&D Program of Ministry of Science and ICT (19-NT-01).

References

1. Dollár P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell* 34(4):743–761
2. Ouyang W, Zeng X, Wang X (2015) Single-pedestrian detection aided by two-pedestrian detection. *IEEE Trans Pattern Anal Mach Intell* 37(9):1875–1889
3. Shen C, Zhao X, Lian X, Zhang F, Kreidieh AR, Liu Z (2019) Multi-receptive field graph convolutional neural networks for pedestrian detection. *IEEE Trans Pattern Anal Mach Intell* 13(9):1319–1328
4. Yoshihashi R, Trinh TT, Kawakami R, You S, Iida M, Naemura T (2018) Pedestrian detection with motion features via two-stream ConvNets. *IPSN Trans Comput Vis Appl* 1–13
5. Qu L, Wang K, Chen L, Gu Y (2016) Free space estimation on nonflat plane based on v -disparity. *IEEE Signal Process Lett* 23(11):1617–1621
6. Lee CH, Lim YC, Kwon S, Lee J (2011) Stereo vision-based vehicle detection using a road feature and disparity histogram. *Opt Eng* 50(2):027004-1-23
7. Labayrade R, Aubert D, Tarel JP (2002) Real time obstacle detection in stereovision on non-flat road geometry through ‘ v -disparity’ representation. In: *Proceedings of the IEEE intelligent vehicle symposium*, 646–651
8. Oniga F, Nedeveschi S (2011) Processing dense stereo data using elevation maps: road surface, traffic isle, and obstacle detection. *IEEE Trans Veh Technol* 59(3):24–36
9. Ess A, Leibe B, Schindler K, Van Gool L (2008) A mobile vision system for robust multi-person tracking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–8
10. Felzenszwalb PF, Huttenlocher DP (2004) Efficient belief propagation for early vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, I-261-I-268

Blockchain-Based Multi-fogcloud Authentication System



Jae Hwan Kwon, Young Kook Kim, Askhat Temir, Kamalkhan Artykbayev, M. Fatih Demirci, and Myung Ho Kim

Abstract Multi-fogcloud, which comprises the cloud system, is a centralized system in which regional fogclouds are managed by a master fogcloud. In a multi-fogcloud, authentication for a regional fogcloud requires sending of an authentication request to the master fogcloud. When regional fogcloud requires excessive authentication, the master fogcloud experiences traffic failures and delays. In this paper to resolve this problem, we suggest a multi-fogcloud authentication system based on blockchain. This system distributes excessive levels of authentication requests through the region fogcloud to resolve problems. It improves authentication speeds by consolidating distributed multi-fogclouds across the blockchain network. We verify the performance of the multi-fogcloud based on blockchain through comparison with the existing multi-fogcloud system.

Keywords Multi-fogcloud · Private blockchain · Hyperledger fabric

J. H. Kwon · Y. K. Kim · M. H. Kim (✉)
Department of Software, Soongsil University, Seoul City, Republic of Korea
e-mail: kmh@ssu.ac.kr

J. H. Kwon
e-mail: jaehwan@soongsil.ac.kr

Y. K. Kim
e-mail: 1101978003@soongsil.ac.kr

A. Temir · K. Artykbayev · M. F. Demirci
Department of Computer Science, Nazarbayev University, Nur-Sultan City, Kazakhstan
e-mail: askhat.temir@nu.edu.kz

K. Artykbayev
e-mail: kamalkhan.artykbayev@nu.edu.kz

M. F. Demirci
e-mail: muhammed.demirci@nu.edu.kz

1 Introduction

In general, multi-fogcloud [1] systems are categorized into a master fogcloud and regional clouds [2]. The master fogcloud features a centralized structure through which it manages and controls multiple regional clouds. As regional fogcloud increases, regional fogclouds make many authentication requests to master fogclouds. When regional fogcloud requests excessive authentication of the master fogcloud, traffic failures [3] and delays [4] may occur. Moreover, the entire multi-fogcloud system can be crashed when the master fogcloud is unable to play the authentication role. Breaking away from the above conventional system in which the master fogcloud exercises centralized control over regional clouds, this paper suggests a decentralized model of multi-fogcloud environment [5] that is enabled through private blockchain [6]. In addition, we compare the performance of the proposed system with that of the existing multi-fogcloud environment.

2 Multi-fogcloud

There are Tacker, Kingbird, and Tricircle in a multi-fogcloud. Among OpenStack-based multi-fogclouds [7], Tacker is an integrated OpenStack project that integrates and manages networks on a cloud infrastructure platform. Based on the MANO (Management and Orchestration) Framework, Tacker uses VNF (Virtualization Network Functions) and supports API. The API can be used to authenticate and monitor individual fogclouds through their heat and keystone. Furthermore, infrastructure configuration facilitates the management of various resources.

Kingbird conducts managed multiple regions management system [8]. Based on OPNFV, it controls Keystone through Kingbird Daemon and Worker, as well as Kingbird API. It also provides resource operation and management features. Kingbird Quotas designates the master fogcloud to provide a centralized quota management service. In addition, it monitors the IP and MAC addresses of sub-fogclouds, manages SSH, and synchronizes images by region.

Tricircle uses Tricircle API to provide data overload control load balancing and traffic management services to each fogcloud. Local Neutron plugin is used to enable centralized management of Neutron servers using API. The neutron deployed in each fogcloud is delivered through the Tricircle Central Neutron plugin. In addition, it manages the tenant IP and MAC addresses and provides collision control over multi-cloud configuration instances.

Table 1 Types of fogcloud variables

Variable name	Description
<i>FOG_IDENTITY</i>	Map form initialization of fogcloud information
<i>X</i>	Transaction value
<i>ERR</i>	Error detection
<i>M</i>	Map for adding or renewing fogclouds
<i>IFGET</i>	MAC address information

3 Design and Implementation

The authentication system between the master fogcloud and the regional fogcloud was designed to construct a multi-fogcloud based on a private blockchain. In this paper used OpenStack in the form of IaaS distributed under the Apache license to configure the fogcloud and private blockchain hyperledger fabric model [9]. Implemented a blockchain-based, multi-fogcloud. Every fogcloud has MAC address information and a unique identifier. The MAC address information of each fogcloud is broadcasted on the blockchain network. MAC addresses are used as unique identifiers for the fogcloud and to authenticate other fogclouds.

3.1 Internal System Structure

Each fogcloud has a hash table and a hash map structure. The fogcloud ID information is unique. The unique values of each fogcloud are used to verify the identity information of the fogcloud. Table 1 shows the actual variable names and descriptions used to store unique values.

3.2 Fogcloud MAC Address Extraction and Verification

To get the new MAC address, the fogcloud information defined in the chaincode is imported. The information is saved in the forms of a hash table and hash map. When the hash table information is returned, the information saved in the fogcloud is printed and displayed as a message. Codes for importing the MAC address value and IP are extracted from the fogcloud. In the main function, `getFogMacAddr()` function is used to receive the printed form of information about the fogcloud. The cloud's information is saved into the transaction in the form of a hash map. Each fogcloud must be able to verify other fogclouds on the blockchain network. *Fog_identity* values hold the MAC addresses of each fogcloud, and are returned in the form of a Boolean value. If an actual value exists, it will be displayed as the MAC address, and if it

does not exist, a “False” message is returned to indicate that there is no information about the fogcloud. To confirm whether the fogcloud exists in the blockchain, the key value is checked, which expresses the fogcloud ID value in the form of hash map as a MAC address. If the MAC address value exists, the “True” value and information about the fogcloud are returned; if it does not exist, the “False” value is returned.

3.3 Fogcloud Authentication Request Codes

To check the fog identity in the new fogcloud, a new MAC address and new IP information are registered on the blockchain network, which is then initialized.

```
func invoke(function string, args []string) ([]byte,
error) {
    var fog_identity map[int]string
    var macaddress string
    var ip string
    var err error
    if fog_identity != nil {
        return nil, err
    }
    //Register Fogcloud MAC Address and IP,
    //Fog Information
    fog_identity := {#macaddress, #fogIP,
#foginfo}
    err = PutState(macaddress, ip)
    if err != nil {
        return nil, er
    }
    return nil, ni
}
```

3.4 Definition of the Chaincodes

The chaincode functions for fogcloud are shown in Table 2. The master fogcloud is the first to initialize the blockchain network. Second, get information about the master fogcloud. Regional fogclouds deliver authentication information to the blockchain network. The master fogcloud and regional fogcloud are broadcasted through the blockchain network. Initialization is executed to activate the blockchain network. In the initialization process, the information about each fogcloud is imported from RocksDB in the form of SQL.

Table 2 Definition of chaincode method

Function	Description
init()	Initialization of chaincodes for the fogcloud
getfog()	Importing IP and MAC addresses of the master fogcloud and the regional fogcloud
invoke()	This is a transaction function that sends information about the fogcloud into other fogclouds. When a new regional fogcloud is added, a transaction function for the new regional fogcloud is generated in other fogclouds as well
query()	Authentication information about the fogcloud is queried

4 Experiments

Open source OpenStack is used to configure a Rocky-version fogcloud environment. Previously, PBFT and chaincodes were configured for transaction agreement with the blockchain network. The blockchain cloud environment that implements a new model and the existing fogcloud environment were compared for various scenarios. The Station is the scope that includes the fogclouds, and contains a master fogcloud and regional fogclouds. The Station can be divided into many entities depending on its physical design. The configured Stations have a master fogcloud that can control and manage their regional fogclouds. In this paper configured Station 1 and Station 2. These stations are separated by a certain physical distance and are connected via a router. There is a distance-vector value between them, which changes depending on the distance connecting the Stations as it goes through the router. Fogclouds that do not go through the router are adjacent fogclouds, while those that do are nonadjacent fogclouds. Depending on the nonadjacent distance-vector values, the distribution time and authentication time of the Tacker [10], Kingbird [11], Tricircle [12] environments, and those of the blockchain-based multi-fogcloud were compared (Table 3).

Table 3 Scenario

Scenario	Explanation
Scenario 1	Regional fogclouds of Station 1 that are adjacent to the master fogcloud of Station 1
Scenario 2	Regional fogclouds of Station 1 that are adjacent to the master fogcloud of Station 1, regional fogclouds of Station 2 that are nonadjacent
Scenario 3	Regional fogclouds of Station 2 that are nonadjacent to the master fogcloud of Station 1, other regional fogclouds of Station 2 that are nonadjacent
Scenario 4-Case 1	Authentication of a new regional fogcloud of Station 2 that is not adjacent to the master fogcloud of Station 1
Scenario 4-Case 2	Authentication of a new regional fogcloud of Station 1 that is adjacent to the master fogcloud of Station 1

To compare the performance with that of existing fogclouds, an average was calculated from a sufficient quantity of data for each scenario design. A total of 100 authentications were executed for each of the existing environments, and the same number of authentications was executed for the proposed method. For Scenario 1, experimental environments for adjacent fogclouds were compared; for Scenario 2, experimental environments for adjacent and nonadjacent fogclouds were compared; for Scenario 3, experimental environments for nonadjacent fogclouds were compared. The authentication times for each of the 100 authentication executions were used to calculate the average authentication time (Fig. 1 and Table 4).

The distance vector is increased according to the scenario. The scenario compared the previous configuration with the blockchain. Comparisons show a difference of 1.05 ms, 0.69 ms, and 6.35 ms, respectively, in Scenario 1. Scenario 2 showed 18.27 ms, 16.76 ms and 23.58 ms respectively. Scenario 3 showed a difference of 52.05, 50.58, and 58.96 ms. Scenario 4 case 1 showed a difference of 51.72, 50.83, 58.56 ms. Scenario 4 case 2 showed a difference of 18.54, 16.6, and 23.52 ms.

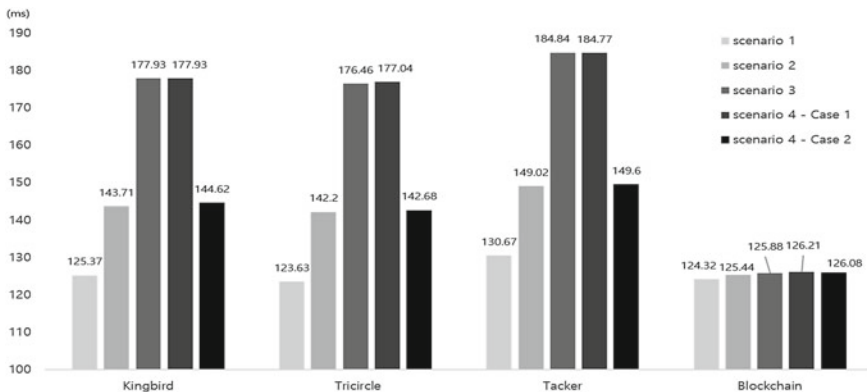


Fig. 1 Average authentication time (out of 100 executions) of Kingbird, Tricircle, Tacker, and blockchain-based environments in each scenario

Table 4 Average for each scenario (ms)

Scenario	Kingbird	Tricircle	Tacker	Blockchain
Scenario 1	125.37	123.63	130.67	124.32
Scenario 2	143.71	142.20	149.02	125.44
Scenario 3	177.93	176.46	184.84	125.88
Scenario 4-Case 1	177.93	177.04	184.77	126.21
Scenario 4-Case 2	144.62	142.68	149.60	126.08

5 Conclusion

This paper proposed a system that can be authenticated through regional fogclouds, without authentication from the master fogcloud. Blockchain-based distributed authentication method can authenticate region fogclouds without a master fogcloud. The performance of the proposed system was experimented in various scenarios. The MAC address and IP information for each fogcloud were saved into the transaction via the chaincodes, to enable authentication of each regional fogcloud. A private blockchain network was used to share the fogcloud information and execute the authentication process. To compare the performance of the proposed system, the authentication times were measured for each scenario. In the comparison of the existing and proposed environments, the distance vector values of the existing fogcloud environments were compared to those of the proposed blockchain-based fogcloud environment to identify the difference in terms of the authentication speed. Comparing the previous configuration with the blockchain configuration shows that the rate of authentication increases as distance vectors increase depending on the scenario. As a result of this experiment, it was confirmed that the blockchain-based authentication method can reduce fogcloud authentication time in a vast majority of the scenarios.

Acknowledgements This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01419) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Dana P (2013) Multi-Cloud: expectations and current approaches. In: Proceedings of the 2013 international workshop on Multi-cloud applications and federated clouds. ACM, 2013, Prague, Czech, pp 1–6
2. Fox A, Griffith R, Joseph A, Katz R (2009) Above the clouds: a Berkeley view of cloud computing. University of California, Berkeley, Dept Electrical Eng and Comput Sciences
3. Krešimir P, Željko H, Hocenski Ž (2010) Cloud computing security issues and challenges. In: The 33rd international convention MIPRO. IEEE, pp 344–349
4. Ekanayake J, Fox G (2009) High performance parallel computing with clouds and cloud technologies. In: International conference on cloud computing. Springer, Heidelberg, pp 20–38
5. Shanhe Y, Zijiang H, Zhengrui Q, Qun L (2015) Fog computing: platform and applications. In: 2015 third IEEE workshop on hot topics in web systems and technologies (HotWeb). IEEE, pp 73–78
6. Zheng Z, Xie S, Dai HN, Wang H (2018) Blockchain challenges and opportunities: a survey. *Int J Web Grid Serv* 14(4):352–375
7. Shi W, Dustdar S (2016) The promise of edge computing. *2016 Computer* 49(5):78–81. IEEE
8. Maël K (2017) Network expansion in OpenStack cloud federations. In: 2017 European conference on networks and communications (EuCNC). IEEE, pp 1–5
9. Cachin C (2016) Architecture of the hyperledger blockchain fabric. In: Workshop on distributed cryptocurrencies and consensus ledgers, vol 310. Switzerland, pp 1–4
10. Openstack project Tacker, <https://wiki.openstack.org/wiki/Tacker>

11. Openstack project Kingbird, <https://wiki.openstack.org/wiki/kingbird>
12. Openstack project Tricircle, <https://wiki.openstack.org/wiki/Tricircle>

Activity-Recognition Model for Violence Behavior Using LSTM



Svetlana Kim, Hyejeong Nam, Hyunho Park, Yong-Tae Lee,
and Yongik Yoon

Abstract Among many dangerous situations, the number of cases of violence has been growing recently. However, there is currently no research to recognize conditions such as assault. Therefore, this paper presents a VR (Violence-Recognition) model for recognition activity using LSTM. The VR model develops algorithms that can detect dangerous situations through processing and analysis of sensing data. Also, to improve accuracy by using the FFT algorithm for processing digital signals in combination with LSTM.

Keywords Smartphone · Smartwatch · Fusion sensing · Abnormal detection · LSTM

1 Introduction

With the development of sensing methods on users' smart devices, more and more various types of data are becoming available for daily research of user activity. The data collected from sensors embedded in smartphones helps identify the behavior

S. Kim · H. Nam · Y. Yoon (✉)

Department of IT Engineering, Sookmyung Women's University, 100, Chungpa-ro 47 gil1,
Yongsan-gu, Seoul 04310, South Korea

e-mail: yyiyeon@sookmyung.ac.kr

S. Kim

e-mail: xatyna@sookmyung.ac.kr

H. Nam

e-mail: nhj93@sookmyung.ac.kr

H. Park · Y.-T. Lee

Smart Media Research Group, Electronics and Telecommunications Research Institute, Daejeon,
South Korea

e-mail: hyunhopark@etri.re.kr

Y.-T. Lee

e-mail: ytleee@etri.re.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_75

information on the users and can highly contribute to the improvement of the HAR (Human Activity Recognition) technology. The HAR referred to as a technology that recognizes user behavior related to user motion movements. With the sensors data can use much useful of human safety protected systems in different domains such as human fall detection [1], personal mood [2] and social behavior [3], healthcare [4], monitoring traffic condition and pedestrian detection [5]. In particular, information capable of recognizing a user's situation is used not only for personal safety but also to provide customized service according to the current case. Most of these studies use more than one sensor and fusion their raw data to obtain meaningful information, which may be more than data from a single-source. Recognized behavioral awareness as one of the key technologies that could bring next-generation smartphones and wearable to the market. Among these dangerous situations, cases of violence have been on the rise recently. Against this backdrop safety concerns have made smart devices an important role in protecting itself. Recently studies have appeared that data collected through acceleration sensors in devices are used to predict the user's behavior or movement [6, 7].

As far as we know, this study on the recognition of violence situation using LSTM (Long Short Term Memory) is the first study. Therefore, this paper presents a VR (Violence-Recognition) model for recognition activity. By analyzing smart device sensors, we extract assault-related features. LSTM is used for classification and FFT (Fast Fourier Transform) is used for accuracy. Another method of classification will be used to compare the performance of classification techniques.

2 Related Work

2.1 Anomaly Detection

Generally the anomaly detection is a technique that detects outlier that show different patterns in the collected data. Anomaly detection algorithm in patterns exists in a variety of ways and the most basic methodologies that correspond to classification are neural network, SVM (Support Vector Machine), and decision tree techniques. Classification model-based anomaly detections use models that are already learned in the course of the test, so the test process is very fast and the results are more accurate, however it is difficult to distinguish anomaly detection from real-time incoming user behavior data. Due to the users' various environments and unpredictable situations, the actual value can be very different to use only the values of fixed parameters at the actual site. Therefore, it can help to achieve better performance in deep learning, which corresponds to unsupervised learning, rather than the classification techniques involved in supervised learning.

2.2 Human-Activity Recognition

Human Activity Recognition technology refers to a technology that uses various sensors to collect and interpret information related to human gestures or motion. In [8] paper shows collecting data from multi-mode sensor such as acceleration, gyro, and altitude sensors were used to learn the attitude and behavior of users through LSTM.

Most paper uses the acceleration sensor that is used as the most influential factor in data measured differently depending on the location of the smartphone. Various data such as acceleration sensors, gyroscope sensors and orientation sensors were used to recognize users' movement such as walking, running, sitting and standing [9, 10].

Therefore, dangerous situation reported in the literature have not been reported under the theme of violence. As a result, it is necessary to study recognition about violence.

3 Implementation

This paper develops algorithms that can detect current users' status through processing and analysis of sensing data from Android-based smartphones and smartwatches, and to classify dangerous situations (Fig. 1).

In this study, 'violence' was defined as an anomaly as a manually designed dangerous situation, because abnormal situations are very infrequent compared to

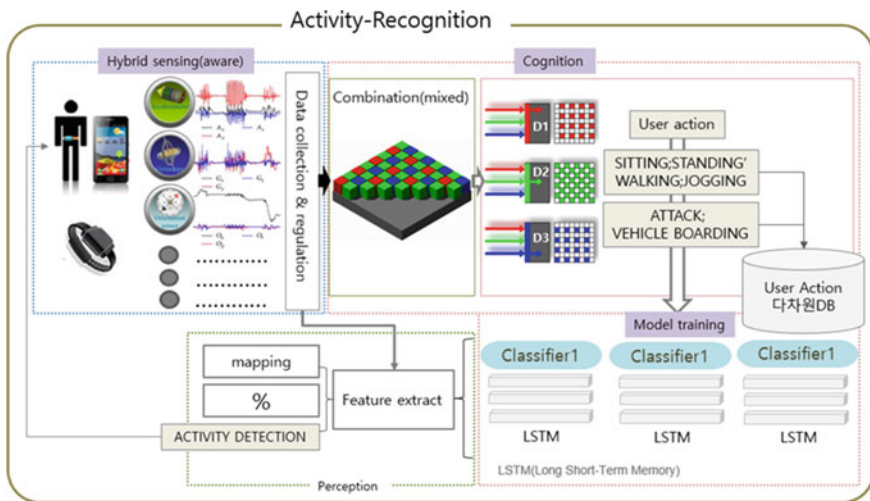


Fig. 1 Activity-recognition

normal situations. The victim's typical behavior was judged to have been that most of the victims hugged their heads when they were attacked.

Modeling four scenarios: (a) Sitting, (b) Sitting attack, (c) Standing and (d) Standing attack. Based on a scenario it collected sensor data with acceleration and gyro sensors through smart phone and smart watch. Compared to everyday life activities such as walking, standing and raising arms through these multimodal sensors, the act of hugging the head when physical forces enter while standing was defined as abnormal values. Therefore in this study, VR would like to conduct a multi-class optimization study to increase the accuracy of risk reasoning.

Learning the normal patterns of sensors uses the LSTM for technique to assess the performance of the abnormal detection according to the sensor patterns. The accuracy was low at 86% when LSTM alone was used, because sensors are digital signals. So there is a lot of noise from vibrations. Accordingly, it was necessary to standardize digital signal through FFT. To model dangerous situations such as assault in Python execution environment, it was combined existing algorithms LSTM and FFT algorithms to increase accuracy.

4 Results

This experiment used smartphones Samsung S8 and LG Sport smartwatch. The data were collected using the Multi log application. It has set four scenarios: (a) Sitting, (b) Sitting attack, (c) Standing and (d) Standing attack, and it has set a cycle of 1 s for Smartphones and 0.3 s for smartwatch.

Figure 2 shows a visualization for extracting the characteristics of the acceleration and gyro sensors from Sitting, Sitting attack, Standing, Standing attack.

Figure 3 shows a visualization of pre-processing data for each action through FFT. The measurements are done at a constant rate of 50 Hz. After filtering out the noise, the signals are cut in fixed-width windows of 2.56 s with an overlap of 1.28 s. Each signal will therefore have $50 \times 2.56 = 128$ samples in total.

Table 1 shows the accuracy of the five classification techniques such as Random Forest, Gradient Boosting Classifier, Logistic Regression, Decision Tree and Nearest Neighbors through detection data obtained from smart phone and smart watch. Random Forest turned out to be the most accurate with smart phone 0.92 and smart watch 0.86, as shown in the results. However, Nearest Neighbors has showed the lowest accuracy of 0.58 on smart phone and Logistic Regression has showed the lowest accuracy of 0.55 on smart watch.

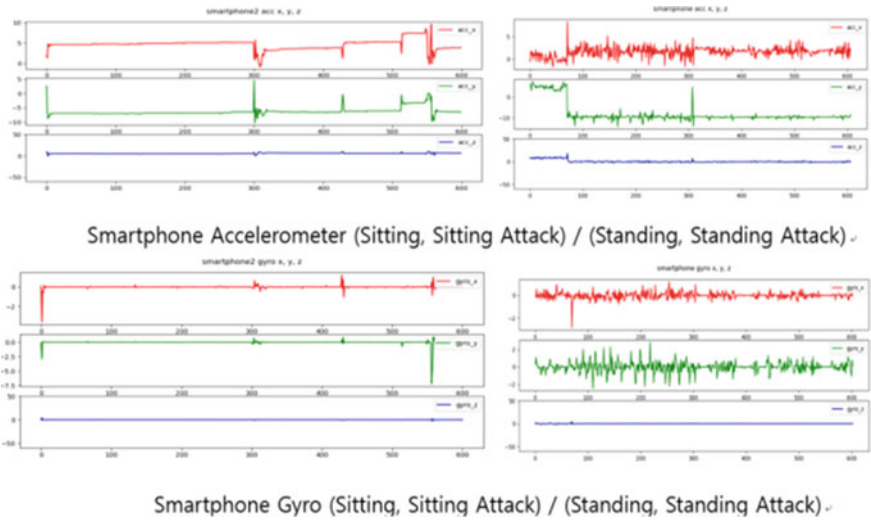


Fig. 2 Smartphone accelerometer and gyro (Sitting, Sitting attack, Standing, Standing attack)

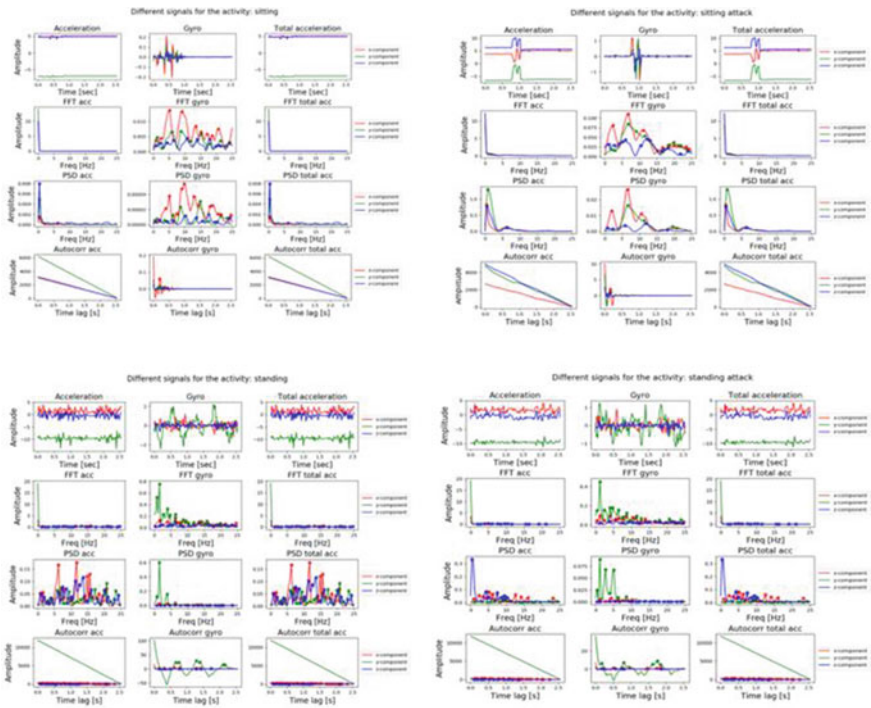


Fig. 3 Visualization of pre-processing data through FFT

Table 1 Accuracy of the classification techniques

Classifier	Smart phone	Smart watch
Random forest	0.92	0.86
Gradient boosting classifier	0.75	0.76
Logistic regression	0.66	0.55
Decision tree	0.66	0.72
Nearest neighbors	0.58	0.76

5 Conclusion

In this paper, acceleration and gyro sensor data were collected through smartphones and smartwatches in order to classify the dangerous situation of violence. By using LSTM to learn the normal pattern of sensors and categorize the dangerous situation according to the difference between forecast results and actual values. However, the accuracy was low at 86% when LSTM alone was used, because sensors are digital signals. So there is a lot of noise from vibrations. Therefore, to improve accuracy by using the FFT algorithm for processing digital signals in combination with LSTM. Through this, high accuracy could be confirmed. For future work, the study is scheduled to use technique of violence classification by combine with video data.

Acknowledgements This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00336, Platform Development of Multi-log based Multi-Modal Data Convergence Analysis and Situational Response).

References

1. Aziza O, Parkc EJ, Morid G, Robinovitch SN (2014) Distinguishing the causes of falls in humans using an array of wearable tri-axial accelerometers. *Gait Posture* 39:506–512
2. Bogomolov A, Lepri B, Pianesi F (2013) Happiness recognition from mobile phone data. In: *BioMedCom* 2013
3. Chittaranjan G, Blom J, Gatica-Perez D (2013) Mining large-scale smartphone data for personality studies. *Pers Ubiquitous Comput* 17(3):433–450
4. Pierleoni P, Pernini L, Belli A, Palma L (2014) An android-based heart monitoring system for the elderly and for patients with heart disease. *Int J Telemed Appl* 11
5. Geronimo D, Lopez AM, Sappa AD, Graf T (2010) Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans Pattern Anal Mach Intell* 32(7):1239–1258
6. Lee S, Mase K (2002) Activity and location recognition using wearable sensors. *IEEE Pervasive Comput* 1:24–32
7. Randell C, Muller H (2000) Context awareness by analysing accelerometer data. In: *The fourth international symposium on wearable computers*, pp 175–176
8. Jahangiri A, Rakha HA (2015) Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE Trans Intell Transp Syst* 16(5):2406–2417

9. Anjum A, Ilyas MU (2013) Activity recognition using smartphone sensors. In: 2013 IEEE consumer communications and networking conference (CCNC), pp 914–919
10. Martín H, Bernardos AM, Iglesias J, Casar JR (2013) Activity logging using lightweight classification techniques in mobile devices. *Pers Ubiquitous Comput* 17(4):675–695
11. Lopez-Cuevas A, Medina-Perez MA, Monroy R, Rez-Marquez JER, Luis A (2018) FiToViz: a visualisation approach for real-time risk situation awareness. *IEEE Trans Affect Comput*, pp 372–373
12. Wu F, Zhao H, Zhao Y, Zhong H (2015) Development of a wearable-sensor-based fall detection system. *Int J Telemed Appl*, Art. no. 2
13. Hengduo L, Jun L, Yuan G, Yirui W (2017) Multi-glimpse LSTM with color-depth feature fusion for human detection. In: IEEE international conference on image processing (ICIP)

Static Analysis for Malware Detection with Tensorflow and GPU



Jueun Jeon, Juho Kim, Sunyong Jeon, Sungmin Lee, and Young-Sik Jeong

Abstract With the advent of malware generation toolkits that automatically generate malware, anyone without a professional skill can easily generate malware. As a result, the number of new/modified malware samples is rapidly increasing. The malware generated in this way attacks vulnerabilities, such as PCs and mobile devices without security patch, causing damages involving malicious actions, such as personal information leakage, theft of authorized certificates, and cryptocurrency mining. To solve this problem, most security companies use the signature-based malware detection technique to detect malware, in which the signatures of known malware and files suspected to be malware are compared before detecting malware. However, the signature-based malware detection technique has a limitation in that it is not efficient for detecting new/modified malware which is generated rapidly. Recently, research is underway to utilize deep learning technology for detecting new/modified malware. In this study, we propose a SAT scheme that can detect not only known malware but also new/modified malware more quickly and accurately, thereby reducing malware-induced damages to PCs and mobile devices. The SAT scheme employs an open source library called Tensorflow in the GPU environment to learn malware signatures and then to statically analyze malware.

Keywords Malware analysis · Malware detection · Static analysis · Deep learning · Signature

J. Jeon · J. Kim · S. Jeon · S. Lee · Y.-S. Jeong (✉)

Department of Multimedia Engineering, Dongguk University, Seoul, Republic of Korea

e-mail: ysjeong@dongguk.edu

J. Jeon

e-mail: jry02107@dongguk.edu

J. Kim

e-mail: 2015112624@dongguk.edu

S. Jeon

e-mail: sunyongj1004@dongguk.edu

S. Lee

e-mail: bearbear11@dongguk.edu

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_76

1 Introduction

About two billion malware attacks occurred in 2018 alone. Since the introduction of automated malware toolkits, about 340,000 new types of malware are detected every day. For rapidly growing new/modified malware, the spreading method to other PCs and mobile devices as well as the symptoms of infected devices are gradually becoming more complicated and intelligent. As a result, PCs and mobile devices infected with such malware experience various hacking-related damages, such as personal and confidential information leakage, cryptocurrency mining, and spam mailing. In the case of recently detected Vidar malware, it is installed in the device by taking advantage of the vulnerability of the Internet Explorer browser where security patch for the vulnerability was not applied, and then infects and spreads the malware by exploiting normal advertisement services [1–5].

In order to protect the user’s PCs and mobile devices from such malware, a variety of techniques for analyzing malware have been investigated. Malware analyzing techniques can be divided into the following three types as shown in Fig. 1. First, the signature-based malware detection technique detects malware by storing signatures of previously detected malware in a database and then comparing them with signatures of files suspected to be malware. In the heuristics-based malware detection technique, which is also called behavior-based malware detection technique, if a certain degree of match is found between specific parts of previously detected malware and a file suspected to be malware, the file is determined to be malware. The specification-based malware detection technique is one type of the heuristics-based malware detection technique. The specification-based malware detection technique is not a method for analyzing the signature of malware, but a method for detecting malware by detecting deviations between the program specification and the

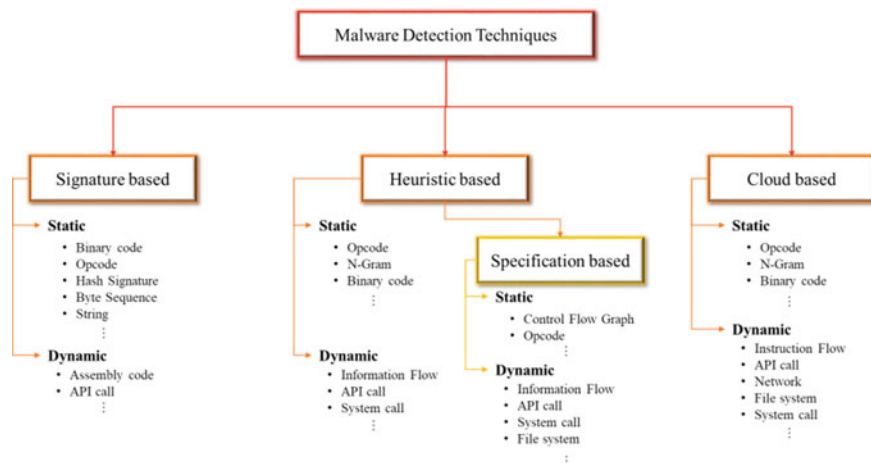


Fig. 1 Various techniques for identifying and detecting malware

program behavior. Lastly, the cloud-based malware detection technique is a method for detecting malware by transferring the file suspected to be malware to a cloud server to analyze the signatures of the malware within the cloud server, and then sending the malware detection result to the client. Most security companies rely on the signature-based malware detection technique to detect malware which attacks and damages PCs and mobile devices and to analyze and identify malware [1–3, 6, 7].

However, while this signature-based malware detection technique can detect known malware accurately, its detection result may not be as accurate for new/modified malware, where part of the malware has been modified or packaged. To solve this problem, many studies have been conducted to apply deep learning to the signature-based malware detection technique to detect malware [1, 3–6, 8].

In an effort to keep the alert level high against the threats of both well-known existing malware and new/modified malware, in this paper, we propose the static analysis for malware detection with Tensorflow (SAT) scheme, which can detect malware quickly and thus prevent it from spreading to other PCs and mobile devices. This SAT scheme employs the Long Short Term Memory (LSTM) model through the Tensorflow library to perform a static analysis of known malware and new/modified malware using the signature-based malware detection technique [9]. The SAT scheme proposed in this study is intended to show the most efficient overall performance with both the speed and accuracy of malware detection being equally considered.

2 Related Works

In order to detect new/modified malware as well as known malware, various studies have been conducted to statically analyze malware using deep learning. Static analysis is defined as a method of analyzing the code and binary file information of malware without directly executing it. In static analysis, opcodes are mainly used as the key signatures for detecting malware. In terms of the opcode, although the opcode itself is important, the sequence between the opcodes is also considered important. For this reason, many studies have been conducted to detect malware by utilizing the LSTM model, in which the learning process is based on the sequence of text strings [10].

As a method to detect IoT malware that threatens to compromise IoT devices used in diverse industries, HaddadPajouh et al. [3] proposed a strategy to build a detection model by extracting the opcodes from decompiled malware and then having them learned by the LSTM. To train the detection model, they used an IoT application data set consisting of 281 malware and 270 non-malware files. In addition, they constructed three different LSTM models to evaluate the detection models which had been trained based on the data set of 100 new malware files. They found that the LSTM model consisting of two hidden layers showed the highest accuracy in detecting new malware compared to the other LSTM models.

Kang et al. [4] created a vector with 1369 dimensions using the one-hot encoding method to classify malware files according to their types and features. After reducing

Table 1 Comparison between the SAT scheme and previous studies

Related works	Feature	Number of branches classified	Performance evaluation factors considered	Target environment
A deep recurrent neural network based approach for internet of things malware threat hunting	Opcode	2 (Benign, Malware)	Accuracy	ARM based IoT devices
Long short-term memory-based malware classification method for information security	Opcode, API call	9 (Malware family)	Accuracy	Window
Our proposed scheme	Opcode, API call	9 (Malware family)	Execution time, accuracy	Window

the number of dimensions from 1369 to 300 using CBoW, which is one of the word2vec technique, they proposed a model that trains an LSTM model consisting of two hidden layers with 128 dimensions to detect malware. For the training of the malware detection model, they used the malware data set published by Microsoft in the Microsoft Malware Classification Challenge (BIG 2015), which is composed of opcodes and API calls extracted through the static analysis of the file contents and characteristics of malware [11]. The performance assessment on the opcode-embedding method of the proposed model led to the conclusion that malware can be detected more quickly and accurately when CBOW, one of the word2vec models, was used compared to the one-hot encoding method.

Table 1 shows the comparison between the SAT scheme proposed in this paper and previous studies.

3 Scheme of SAT

In order to establish a malware detection model with efficient performance in terms of the speed and accuracy of detection and classification during the detection and classification process of malware with various characteristics, in this paper, we propose the SAT scheme that performs static analysis of malware in the GPU environment utilizing the Tensorflow library.

The SAT scheme consists of three phases: data processing, pre-processing, and learning phases, as shown in Fig. 2, where malware learning and detection are processed based on this schematic. In the first data processing phase, for the training of

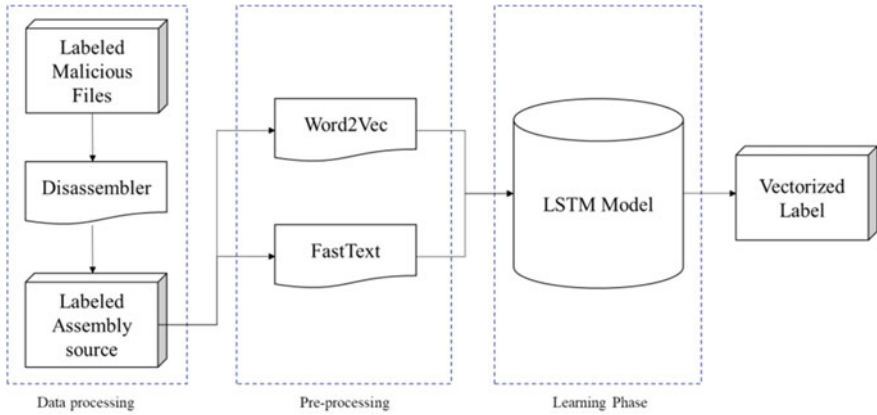


Fig. 2 Overall schematic of the SAT scheme for malware detection and classification

malware signatures, known malware files are labeled and statically analyzed through a disassembler to extract their opcodes and API calls. In the pre-processing phase, word embedding is performed using word2vec or fasttext to apply the extracted opcode sequence data to the LSTM model. In the last learning phase, the malware is classified according to the characteristics of the malware based on the signatures of the vectorized malware, which are opcodes and API calls, followed by malware learning and detection.

3.1 Data Processing

In the data processing phase, the malware, which has been classified into one of the nine kinds of malware, is labeled to extract its opcodes and API calls, which are the signatures of the malware, and through a disassembler process, the malware is converted from a machine code form to an assembly language code form to statically analyze the file information and the code contents of the malware.

3.2 Pre-processing

In the pre-processing phase, a word embedding method, through which opcodes and API calls in the form of natural language are converted to numbers for computers to understand and efficiently process them, is applied so that the learning process is carried out in the LSTM model based on the opcodes and API calls extracted in the static analysis during the previous data processing phase. The SAT scheme proposed in this paper employs a word embedding method of word2vec or fasttext, whichever

shows optimal performance, in the pre-processing phase. The word2vec technique was developed by Google in 2013 and is one of the techniques for vectorizing key words by analyzing surrounding assertions [12]. Word2vec consists of the CBoW technique that predicts words based on context and the Skip-gram technique that predicts context based on one word. Fasttext is a technique developed by Facebook that considers many subwords to exist in a single word and vectorizes the word in consideration of the subwords [13].

3.3 Learning Phase

Lastly, the learning phase, where malware is detected by learning and classifying the signatures of malware according to their characteristics, is composed as shown in Fig. 3. The opcodes and API calls vectorized through the pre-processing process are assigned to two LSTM layers composed of 128 cells, which are hidden layers, in order to classify them into one of the nine kinds of malware which have been classified by the characteristics of malware. Here, the nine kinds of malware includes Ramnitm, Lollipop, Kelihos_ver3, Vundo, Simda, Tracur, Kelihos_ver1, Obfuscator.ACY, and Gatak. Malware is learned through Softmax layer and Adam Optimizer, and the new malware is tested based on the model established in this way.

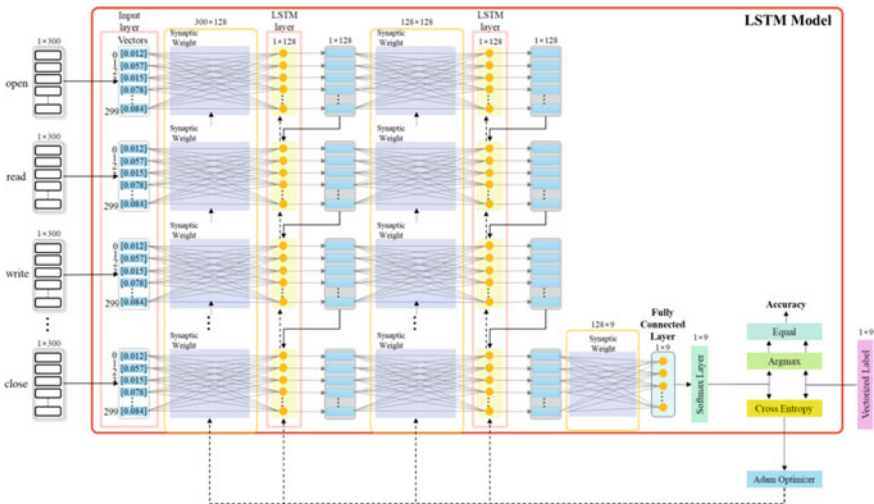


Fig. 3 The learning phase in the SAT scheme for detecting and classifying malware

4 Experience

The SAT scheme proposed in this paper was tested under the environments with the configurations of CPU with eight AMD FX-8370E Eight-Core Processors and of GPU of Quadro P4000 with 32.9 GB memory using the data set published by Microsoft in BIG 2015 [11]. The main purpose of the experiment conducted for the SAT scheme is to determine whether to use word2vec or fasttext as the word embedding method in the static analysis of malware to obtain the most optimized results for the SAT scheme. The input data size, window size, hidden layer number, and cell number were used as parameters for both word2vec and fasttext. Additionally, the embedding method was used as a parameter for word2vec for analysis, while the n-gram range was used as a parameter for fasttext for analysis. Input data refers to the data to be entered in the word-embedding method after extracting API calls and opcodes from the assembly code and window refers to the range of data analysis in the word-embedding method. In addition, cell refers to the size of the hidden layer and the hidden layer refers to the step of input data processing within the LSTM model. The initial values used in the experiment were as follows: 300 for the input data, 5 for the window size, 128 for the number of cells, and 2 for the hidden layer.

Tables 2 and 3 show the accuracy of malware detection when the word2vec model or the fasttext model was used as the word embedding method. The accuracy of malware detection under the various conditions ranged from 96.61 to 97.60%. In addition, a higher accuracy of 97.60% was observed in detecting malware when fasttext was used compared to when word2vec was used.

The word2vec technique showed a malware detection accuracy of 97.59%, and the fasttext technique showed a malware detection accuracy of

Table 2 Detection accuracy when using the word2vec model as the word embedding method

Type	Input data	Window size	Cell	Hidden layer	Embedding method	Accuracy (%)
Word2vec	200	5	128	2	CBoW	97.57
	300	5	128	2	CBoW	97.59
	400	5	128	2	CBoW	97.59
	300	3	128	2	CBoW	97.59
	300	5	128	2	CBoW	97.59
	300	7	128	2	CBoW	97.59
	300	5	64	2	CBoW	97.54
	300	5	128	2	CBoW	97.59
	300	5	256	2	CBoW	97.59
	300	5	128	2	CBoW	97.59
	300	5	128	3	CBoW	97.19
	300	5	128	2	CBoW	97.59
	300	5	128	2	Skip-gram	97.59

Table 3 Detection accuracy when using the fasttext model as the word embedding method

Type	Input data	Window size	Cell	Hidden layer	N-gram	Accuracy (%)
Fasttext	100	5	128	2	3-6	97.59
	200	5	128	2	3-6	97.59
	300	5	128	2	3-6	97.60
	300	3	128	2	3-6	97.60
	300	5	128	2	3-6	97.60
	300	7	128	2	3-6	97.60
	300	5	64	2	3-6	97.51
	300	5	128	2	3-6	97.60
	300	5	256	2	3-6	97.59
	300	5	128	2	3-6	97.59
	300	5	128	3	3-6	96.61
	300	5	128	2	2-5	96.60
	300	5	128	2	3-6	97.60
	300	5	128	2	4-7	97.60

Table 4 Execution time when word2vec or fasttext was used in the pre-processing phase

Word2vec	12 min
Fasttext	41 min

Table 5 Execution time under the CPU and GPU environments in the learning phase

		Device	
		CPU	GPU
Word embedding	Word2Vec	42 h 20 m 30 s	8 h 12 m 23 s
	Fasttext	43 h 45 m 10 s	8 h 16 m 2 s

97.60%. Tables 4 and 5 show the analysis results of the execution time when word2vec and fasttext were executed under the optimal conditions in the pre-processing and learning phases. Table 4 shows the execution time when word2vec and fasttext were used as the word embedding method in the pre-processing phase. Table 5 shows the execution time when word2vec and fasttext were executed in the CPU and GPU environments in the learning phase.

5 Conclusion

With the advent of toolkits that can automatically generate malware, anyone can create new/modified malware without being an expert. As a result, the number of malware samples is rapidly increasing, and security companies spend a lot of time in analyzing the characteristics of new/modified malware and generating signatures of malware to help detect malware. To solve this problem, various malware detection techniques have emerged, and many relevant studies have been conducted. However, catching up with the generation rate of new/modified malware is impossible, which makes it difficult to analyze and detect the characteristics of new/modified malware. For this reason, deep learning has been applied to help detect malware, and the related research has begun.

In static analysis, opcodes and API calls are considered as the signatures of malware, which are the feature elements of malware. For these opcodes, a single word itself is not meaningful, but the surrounding words are rather meaningful. Therefore, the LSTM model has been mainly used to analyze malware using opcodes. However, previous studies utilizing the LSTM model focused solely on the accuracy of malware detection.

Therefore, in this paper, we proposed the SAT scheme that statically analyzes malware in the GPU environment to detect malware quickly and accurately using the LSTM model. The SAT scheme consists of three phases: data processing, pre-processing, and learning phases. The results of the experiments on the malware detection time and accuracy, which were performed based on the SAT scheme, indicated that fasttext was more efficient in terms of accuracy than word2vec, but when fasttext was used in the pre-processing phase, an inefficient execution time was observed. However, when the learning phase, which is the time when malware is classified and learned in the LSTM model, was compared, the execution time under the GPU environment was about three times more efficient than that under the CPU environment. This suggests that when fasttext is used, the problem of inefficient execution time in the pre-processing phase can be compensated.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00644, Linux Malware Dynamic Detection & Protection Solution on Embedded Device).

References

1. Souri A, Hosseini R (2018) A state-of-the-art survey of malware detection approaches using data mining techniques. *Human-Centric Comput Inf Sci* 8:1–22
2. Keegan N, Ji S-Y, Chaudhary A, Concolato C, Yu B, Jeong DH (2016) A survey of cloud-based network intrusion detection analysis. *Human-Centric Comput Inf Sci* 6:1–16
3. HaddadPajouh H, Dehghantanha A, Khayami R, Choo K-KR (2018) A deep recurrent neural network based approach for internet of things malware threat hunting. *Future Generat Comput Syst* 85:88–96

4. Kang J, Jang S, Li S, Jeong Y-S, Sung Y (2019) Long short-term memory-based Malware classification method for information security. *Comput Electr Eng* 77:366–375
5. Choi S-Y, Lim CG, Kim Y-M (2019) Automated link tracing for classification of malicious websites in malware distribution networks. *J Inf Process Syst* 15:100–115
6. Daoud WB, Obaidat MS, Meddeb-Makhlouf A, Zarai F, Hsiao K-F (2019) TACRM: trust access control and resource management mechanism in fog computing. *Human-Centric Comput Inf Sci* 9:1–18
7. Belaoued M, Mazouzi S (2016) A Chi-square-based decision for real-time malware detection using PE-file features. *J Inf Process Syst* 12:644–660
8. Nagpal B, Chauhan N, Singh N (2017) A survey on the detection of SQL injection attacks and their countermeasures. *J Inf Process Syst* 13:689–702
9. Tensorflow. <https://www.tensorflow.org/?hl=ko>
10. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neu Computat* 9:1735–1780
11. Microsoft Malware Classification Challenge (BIG 2015). <https://www.kaggle.com/c/malware-classification>
12. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: *ICLR 2013, International conference on learning representations. Conference Track Proceedings, Arizona*, pp 1–12
13. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2017) Advances in pre-training distributed word representations. In: *The Eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, pp 52–55

IoT Malware Dynamic Analysis Scheme Using the CNN Model



Jueun Jeon, Seungyeon Baek, Minhoo Kim, Inho Go, and Young-Sik Jeong

Abstract Recently, Internet of Things (IoT) technologies have been fused with next-generation technologies such as 5G and deep learning and used in diverse fields such as smart homes, smart cars, and smart appliances. As the demand for IoT devices increases, security threats targeting IoT devices, IoT infrastructure, and IoT application programs have also been increasing. Diverse studies on IoT malware detection have been conducted to protect IoT devices particularly from IoT malware among the security threats. However, existing studies can only accurately detect known IoT malware, not new and variant IoT malware. In this study, the malware dynamic analysis (MALDA) scheme that accurately detects new and variant malware that threatens IoT devices quickly is proposed to reduce the damage caused to IoT devices. The MALDA scheme dynamically analyzes IoT malware in nested cloud environments by training the behavioral features of IoT malware based on the Convolutional Neural Network (CNN) model.

Keywords Internet of things · Malware · Malware detection · Dynamic analysis · Deep learning

J. Jeon · S. Baek · M. Kim · I. Go · Y.-S. Jeong (✉)
Department of Multimedia Engineering, Dongguk University, Seoul, Republic of Korea
e-mail: ysjeong@dongguk.edu

J. Jeon
e-mail: jry02107@dongguk.edu

S. Baek
e-mail: tmddusdls12@dongguk.edu

M. Kim
e-mail: aceed7592@dongguk.edu

I. Go
e-mail: akdlwbek@dongguk.edu

1 Introduction

In the IoT environment, attack surfaces can be infinitely expanded because all devices, including things, spaces, and data, are connected to each other. Some manufacturers are mass producing IoT devices that are vulnerable in security to respond to the rapidly changing market. IoT devices have become highly likely to be the major targets of malware producers because of their vulnerability. According to Kaspersky Lab's IoT report, the number of malware samples that threatened IoT devices was 32,614 in 2017, and it increased by four times to 121,588 in 2018 as more than 120,000 variants of malware that attacked IoT devices were found. When IoT devices are infected with malware, not only are they abused for illegal activities, such as cryptojacking, DDoS attacks, and botnet activity, among others, but users' personal information collected in them is also extorted [1–5].

Therefore, studies have been conducted to detect IoT malware to protect IoT devices from security threats. Malware detection techniques used in the studies include signature-based, heuristics-based, specification-based, and cloud-based ones, and each one includes static analysis and dynamic analysis as the malware analysis methods [1, 4, 6–8].

However, detecting IoT malware that propagates quickly to other devices in real time from IoT devices with limited resources is difficult. Moreover, the number of samples of new and varying IoT malware is rapidly increasing because of the emergence of automated malware generation tool kits, and the methods for attacking IoT devices are also changing intelligently [9–11].

Therefore, this study proposes a scheme that trains the behavioral features of malware based on the Convolutional Neural Network (CNN) model and performs a dynamic analysis of malware in nested cloud-based virtual machine environments to determine whether IoT malware has been detected or not in real time.

2 Related Works

Malware analysis methods are largely divided into static analysis and dynamic analysis. Static analysis analyzes malware by focusing on the analyses of binary file information and codes, and dynamic analysis examines malware samples to determine whether malware exists based on the extracted action data. However, static analysis can hardly detect obfuscated malware and cannot identify all functions of malware.

Therefore, diverse studies have been conducted using dynamic analysis techniques to identify the overall functions of malware that threatens IoT devices with limited resources and to accurately detect the relevant malware.

HaddadPajouh et al. [9] proposed a method for detecting IoT malware that damages IoT devices using the long short-term memory model, which is one of the recurrent neural networks. The data used to train the detection model are an ARM

processor-based IoT application consisting of 281 pieces of malware and 270 non-malware applications. Opcodes are extracted from the data set based on the features of the IoT malware to carry out learning and analysis. However, as the proposed model analyzes and detects IoT malware in the Raspberry Pi, which can be said to be an IoT device, the problem of the speed and accuracy of IoT malware detection varies with the specification of the IoT device.

Wu [10] proposed a system for detecting android malware through the K-nearest neighbor (KNN) model in the android platform environment. The proposed system extracts API calls as features from a data set consisting of 1,160 non-malware files and 1,050 malware files, preprocesses the API calls into the form of a call graph, and then applies the call graph to the KNN model to train and classify android malware.

3 MALDA Scheme

This study proposes the MALDA scheme that dynamically analyzes malware in a cloud-based environment to detect IoT malware that attacks IoT devices with limited resources in real time.

MALDA consists of a nested virtual environment. As most IoT devices are driven by an embedded Linux based on an ARM processor, a nested environment is constructed to analyze IoT malware that operates in the ARM process.

The MALDA scheme detects IoT malware in a nested cloud virtual environment through a malware detection process as shown in Fig. 1. The malware detection process in the MALDA scheme is identical to the general malware detection process up to the feature extraction stage. After extracting the behavioral features, the malware detection process in the MALDA scheme undergoes preprocessing to conduct the imagification of the behavioral features as input values into the CNN

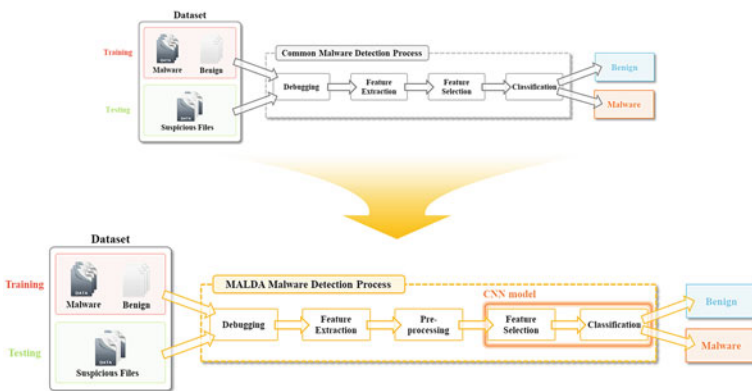


Fig. 1 General malware detection process and the malware detection process in the MALDA scheme

model; this step is called the classification stage. The behavioral features made into images are used as input values in the CNN model, and the CNN model trains the behaviors of malware through the feature selection and classification processes and classifies the behaviors [8, 11].

3.1 Debugging

In the debugging stage, files are executed in a nested virtual environment to directly check the flow of the code and memory status in the data sets for the collected IoT malware and non-malware files. The debugging of files executed in the nested virtual environment is conducted remotely using an analysis tool called IDA Pro, which generates assembly language codes from binary files.

3.2 Feature Extraction

In the feature extraction step, the feature that can be regarded as the signature of IoT malware is extracted based on the system call, process scheduling, and network information generated after analyzing the internal structure of actions and files through the debugging process. The extracted features are largely classified into five categories, namely, memory, network, process, system call, and virtual file system, and the features are stored in the Excel file format.

3.3 Preprocessing

The behavior log data used as input values in the preprocessing stage are divided into three types: unordered log data, ordered log data, and semantic data. These three types of log data undergo a three-stage preprocessing to be used in the CNN model. All the strings in the log data are given an id and then made into integers. The values that become integers are rescaled to values between 0 and 255 for imagification. The matrices are then generated in diverse sizes according to the log data. Finally, in order to prevent matrices with various sizes according to the log data, the image was unified to on size through the resizing technique. As a result, the unordered log data generate images in the R channel, the ordered log data in the G channel, and the semantic log data in the B channel. The images are combined into one image, which is stored.

3.4 Feature Selection and Classification

In the MALDA scheme, the feature selection stage, in which the representative behavioral features are selected from the IoT malware features made into images, and the classification stage, in which such behavioral features are learned and classified into malware and non-malicious files, are integrated into one stage using the CNN model to detect IoT malware. Here, ZFNet is used as the CNN model. ZFNet is an algorithm that won first place in the 2013 ImageNet Large Scale Visual Recognition Challenge competition, and its recognition error rate is 11.2% [12].

The behavioral features are detected from the behavioral feature images generated in the preprocessing stage, and a matrix called a feature map is produced from the behavioral features. To reduce the dimension of the generated feature map, the max pooling technique is applied to extract only the feature with the largest value. When the feature selection stage in the ZFNet model has been completed, the behavioral features of IoT malware are trained according to the feature map to classify the files into malware or non-malicious files.

4 Experiment

The proposed MALDA scheme was experimented for IoT malware detection in the environment configured as follows.

4.1 Dataset and Debugging

A total of 900 malware samples and 1,056 non-malicious files were collected. A total 800 malware samples and 960 non-malicious files were used as training data. A total of 100 malware samples and 96 non-malicious files were used as testing data.

The training and testing data sets were executed for 5 min in a nested virtual environment in the cloud. During this time, remote debugging of the files was performed.

4.2 Feature Extraction

The log data generated in the debugging stage were classified into memory, network, process, system call, and virtual file system, and they were extracted in the form of Excel.

4.3 Preprocessing

Based on the memory, network, process, system call, and virtual file system created in the behavioral feature extraction stage, the images of the R, G, and B channels were created. Each channel was integrated into one image.

4.4 Feature Selection and Classification

The images generated in the preprocessing stage were substituted as input data to learn the behavioral features extracted from the training data set, and a malware detection model of the MALDA scheme was constructed. Based on the established detection model, the test data set was classified into malware or non-malicious files. According to the results, the detection accuracy for IoT malware was 100%.

5 Conclusion

IoT devices that are vulnerable in security are infected because of the rapidly generated new and varying IoT malware, as countermeasures against malware cannot be prepared and the malware is spread to other devices. To solve this problem, many studies have been conducted to analyze and detect IoT malware. However, in existing studies, experts intervened to identify and select the representative behaviors of malware. Thus, accurately detecting the new and varying IoT malware that is rapidly generated is difficult.

This paper proposed a MALDA scheme that dynamically analyzes IoT malware based on the CNN model in a nested virtual environment in the cloud. When an IoT device sends a file suspected to be malware, the MALDA scheme detects the IoT malware in real time through a five-stage IoT malware detection process.

Therefore, preventing the IoT malware from spreading to or infecting different IoT devices connected with each other in a network is possible through the MALDA scheme. Even when IoT malware threatens IoT devices with diverse intelligent attack techniques, it can be accurately detected using the CNN model.

As the training and testing data sets used in the proposed MALDA scheme consisted of variants of Mirai malware that occurs mainly in IoT devices, the malware detection accuracy of the MALDA scheme was shown to be 100%. Therefore, in future studies, diverse types of IoT malware will be collected to further conduct training and tests.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1088383).

References

1. Jeong Y-S, Park JH (2019) IoT and smart city technology: challenges, opportunities, and solutions. *J Inf Process Syst* 15:233–238
2. Choi S-Y, Kim D, Kim Y-M (2016) ELPA: emulation-based linked page map analysis for the detection of drive-by download attacks. *J Inf Process Syst* 12:422–435
3. Nagpal B, Chauhan N, Singh N (2017) A survey on the detection of SQL injection attacks and their countermeasures. *J Inf Process Syst* 13:689–702
4. Choi S-Y, Lim CG, Kim Y-M (2019) Automated link tracing for classification of malicious websites in malware distribution networks. *J Inf Process Syst* 15:100–115
5. Kaspersky. <https://news.kaspersky.co.kr/news2018/09n/180920.htm>
6. Alghamdi TA (2019) Convolutional technique for enhancing security in wireless sensor networks against malicious nodes. *Human-Centric Comput Inf Sci* 9:1–10
7. Keegan N, Ji S-Y, Chaudhary A, Concolato C, Yu B, Jeong DH (2016) A survey of cloud-based network intrusion detection analysis. *Human-Centric Comput Inf Sci* 6:1–16
8. Sourì A, Hosseini R (2018) A state-of-the-art survey of malware detection approaches using data mining techniques. *Human-Centric Comput Inf Sci* 8:1–22
9. HaddadPajouh H, Dehghantanha A, Khayami R, Choo K-KR (2018) A deep recurrent neural network-based approach for internet of things malware threat hunting. *Future Generat Comput Syst* 85:88–96
10. Wu S, Wang P, Li X, Zhang Y (2016) Effective detection of android malware based on the usage of data flow APIs and machine learning. *Inf Softw Technol* 75:17–25
11. Sharmeen S, Huda S, Abawajy JH, Ismail WN, Hassan MM (2018) Malware threats and detection for industrial mobile–IoT networks. *IEEE Access* 6:15941–15957
12. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: 13th European conference on computer vision (ECCV), computer vision—ECCV 2014, Zurich, pp 818–833

A Design of Improvement Method of Central Patch Controlled Security Platform Using Blockchain



Kyong-Tack Song, Shee-Ihn Kim, and Seung-Hee Kim

Abstract The enterprise patch management system is highly dependent upon a central management server and typically suffers a single point of failure. This study suggests a hyperledger fabric blockchain-based distributed patch-management system and verifies its technological feasibility through prototyping. This is the first study to use blockchain technology for a patch-management system. It not only provides a higher level of security and availability but also affords the basis for further review when planning new business discovery of the industry to apply blockchain technology.

Keywords Blockchain · Patch · Hyperledger fabric

1 Introduction

Patches are usually included in new software releases so that product developers can provide tailored solutions and improved security measures [1]. All information-technology (IT)-based companies must apply dedicated plans to perform patch-management activities as part of their enterprise management system. With this, an organization-wide strategy must be established to govern how and when patches are applied to certain systems to manage and mitigate vulnerabilities.

This implies that companies depending more on IT capabilities not only require higher-level software patch strategies, they require stronger security-minded leadership. However, the important issue here involves understanding how systems having

K.-T. Song · S.-I. Kim · S.-H. Kim (✉)

Department of IT Convergence Software Engineering, Korea University of Technology & Education (KOREATECH), 1600, Chungjeol-ro, Byeongcheon-myeon, dongnam-gu, Cheonan, Chungnam 31253, Republic of Korea
e-mail: sh.kim@koreatech.ac.kr

K.-T. Song
e-mail: skt2000@koreatech.ac.kr

S.-I. Kim
e-mail: eftpos@koreatech.ac.kr

cybersecurity vulnerabilities are updated with usable patches. This is because there is no guarantee that the patch will not conflict with other applications running on the network. Sometimes a balance of priorities is required to minimize interruptions of mission critical systems [2]. Nevertheless, patch management is essential to protect software from cybersecurity threats and to lessen the burdens of system administrators. Patch management requires cognizant support from high-level managers, dedicated resources, clearly defined and assigned responsibilities, creation and maintenance of current technology inventories, vulnerability and patch identification, network scanning and monitoring, pre-distribution test patching, and post-distribution scanning and monitoring [3]. This study defines a patch-management system as a software system that helps administrators perform company-wide mass management and control of patches and updates for operating systems (OS) and applications on all servers and user computers on a network [4].

We design and verify the technological validity of a hyperledger fabric blockchain-based distributed patch-management system that can reliably protect clients from a variety of threats. To this end, the blockchain platform uses peer-to-peer (P2P) networking, Proof-of-Work (PoW) validation, and distributed ledger technology that does not depend on a central server [5] and can transparently and safely manage all transactions.

2 Preliminary

2.1 Background Research

Previous studies on patch management mainly included works focusing on improving system stability and reliability and on design and implementation of patch-management system improvements (Table 1).

2.2 Blockchain Consensus Algorithm

In a typical public blockchain, a PoW algorithm [14] is used. This is the consensus algorithm that was used in the initial version of the Bitcoin. Ethereum, which handles electronic transactions, used a consensus algorithm that includes both PoW and proof-of-transaction. PoW uses a reverse function to find a certain hash value. As such, it requires advanced computing power.

The proof-of-stake (PoS) algorithm, introduced by Ethereum, was developed to resolve the problem of electricity waste caused by traditional PoW computing. This method found consensus based on assets possessed by nodes. The problem with PoS is that it can be restricted by a certain person, because it grants the discretion to create blocks to the node possessing the stake. To compensate for this problem, the delegated

Table 1 Overview of previous studies

Author	Ref. #	Overview of previous studies
May et al.	[6]	The game theoretical model to find a balance between the costs and benefits of patch management and to study the strategic interaction between companies and suppliers
Gerace and Cavusoglu	[3]	The importance of core elements in patch-management processes
Kim et al.	[7]	A method of differentiating the same OS by client and performing various patch deployments
Kim et al.	[8]	A web crawler to analyze the structure and characteristics of vendor sites for this purpose
Muhammad and Sinnott	[9]	A trust-oriented policy-centered infrastructure that could overcome many of the problems occurring because of architectural unconditional trust assumptions
Chang et al. and Lee et al.	[10] and [11]	A patch-management process that supported heterogeneous environments and used a process cycle and patch strategy that increased the efficiency of enterprise processes and reduced patch-management risks
Midtrapanon et al.	[12]	The cost down effort of using open source program with customization
Zheng et al.	[13]	The long time checking interval based path management strategy is increasing the security level of VM-base ITS

PoS (DPoS) method [15] was developed. The DPoS method is similar to PoS, but it grants the right to create blocks to the top 101 nodes elected by vote. Hyperledger is a private blockchain, and it has an improved consensus method compared with Bitcoin and Ethereum. Its consensus algorithm uses a process flow that broadly consists of endorsement, orderers, and validations.

3 Central Patch Controlled Security Platform Using Blockchain

3.1 Architecture of Patch-Management System

As shown in Fig. 1, the patch-management service provides patch-file management, user management, patch-status management, and patch-file verification. The deployment service performs patch-file deployment, patch integrity checks, and patch-file verification. The deployment service includes a patch-status distributed database and

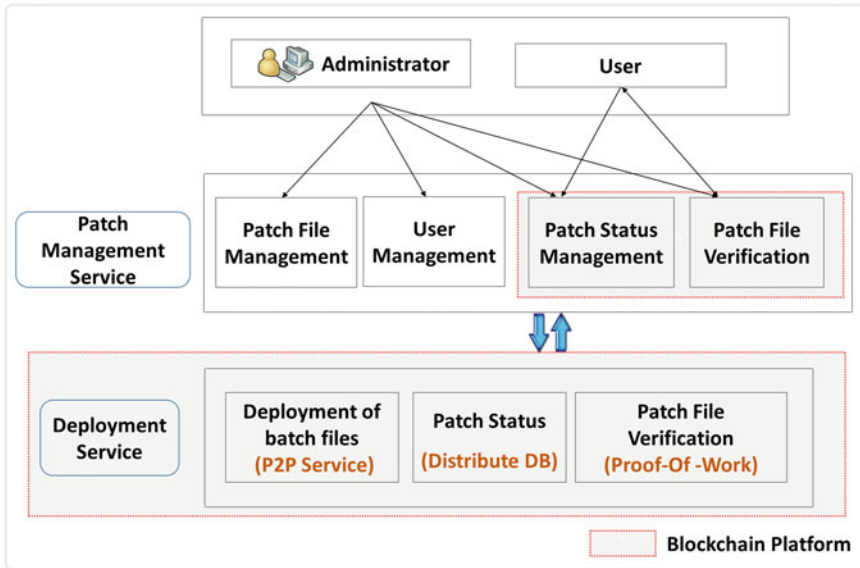


Fig. 1 Architecture of patch-management system

a patch-file verification consensus algorithm for verifying patch-file integrity. The scope of this study includes the portions marked with the red dotted line Fig. 1.

3.2 Concepts of the Patch-Management System

The blockchain-based patch-management system proposed in this study comprises a patch-management service and a blockchain platform-based deployment service, as shown in Fig. 2. First, the service manager registers the patch file with the patch-management server. The patch-management server then stores the registered file and sends it to a blockchain-based deployment service network of service users. The patch-file set sent to the deployment service network is shared to each node (i.e., service user) connected to the blockchain platform.

3.3 Research Procedure

Step 1: Principal Concepts of the Patch-management System.

The detailed process can be examined via the sequence diagram.

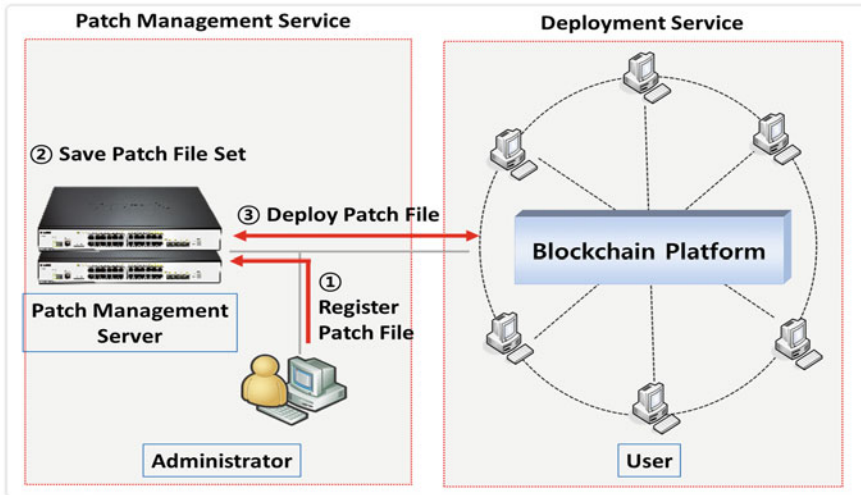


Fig. 2 Conceptual diagram of blockchain-based patch-management service

Step 2: Blockchain Configuration and Block Structure Design.

There are two nodes (i.e., peers) for each organization (Org1 and Org2). Individual nodes store the same blockchain distributed ledger. Each organization performs identity verification using a membership service provider (MSP) with certificate authority.

The blockchain configured in this manner is assigned to a channel according to usage rights. In this configuration, a single channel is set up.

Step 3: Implementation and Experiments.

The patch-management system comprises a private blockchain platform.

We test the feasibility of implementing a blockchain-based patch-management system by confirming that the patch-file set admin uploaded the set to the block chain, that the user downloaded the patch-file set, and that the log record is used to verify downloading. The experiment is divided into three parts as shown in Fig. 3: uploading the patch file set, downloading the patch files and recording logs, and download verification. The source file for the environment settings of the hyperledger fabric blockchain platform is created as a YAML data serialization file.

Step 4: Experiment Results and implication.

The chaincodes are created for the patch-file deployment service. The patch-file set data registered by the administrator include all of these data, used as chaincode in the blockchain system. A connection between the node and the channel is required to execute the features defined by the chaincode. The mutual connection is performed internally by the hyperledger fabric system. The implications of this study are then reviewed.

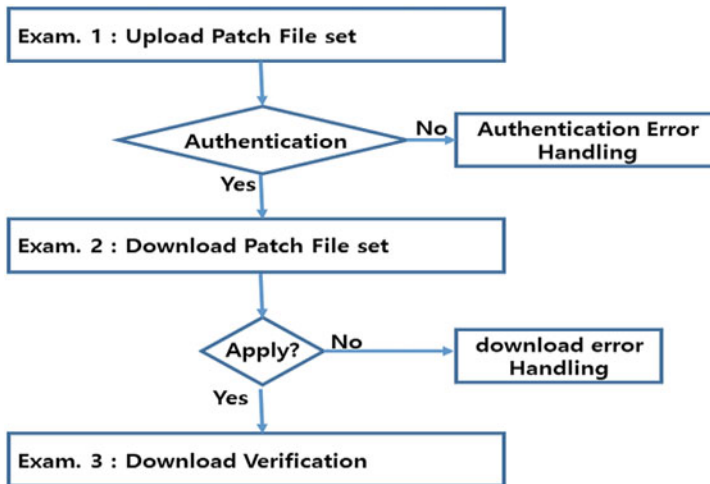


Fig. 3 Test scenario flow

4 Conclusion

This study proposed a hyperledger fabric blockchain-based distributed patch-management system and verified its technological feasibility via prototyping, enabling all participants to be protected from various deployment threats. In the deployment service, patch files were deployed to the agent via P2P. The blockchain's distributed database storage method and a PBFT consensus algorithm technique were used to verify that the patches were executed normally. By doing so, the integrity of the patch application status database is verified.

References

1. Dennis C (2018) Why is patch management necessary? *Netw Secur* (7):9–13
2. Liu S, Kuhn R, Rossman H (2009) Surviving insecure it: effective patch management. *IT Profess* 11(2):49–51
3. Gerace T, Cavusoglu H (2005) The critical elements of patch management. In: Proceedings of the 33rd annual ACM SIGUCCS conference on user services, pp 98–101
4. Lee IY, Lee SY, Moon JS, Lim JI (2008) A study on efficient component in patch management system. In: Proceedings of the Korean Society of Broadcast Engineers Conference. The Korean Institute of Broadcast and Media Engineers
5. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system
6. May P, Ehrlich HC, Steinke T (2006) ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: European conference on parallel processing. Springer, Berlin, Heidelberg, pp 1148–1158
7. Kim YJ, Lee SW, Sohn TS, Moon JS, Seo JT, Yun JB, Park EK (2004) Design the classed patch distribution system framework considering the extension. In: Proceedings of the 2004 Fall The

- Korean Institute of Information Scientists and Engineers, The Korean Institute of Information Scientists and Engineers, vol 31, no 1A, pp 199–201
8. Kim Y, Na S, Kim H, Won Y (2018) Automatic patch information collection system using web Crawler. *J Korea Inst Inf Secur Cryptol* 28(6):1393–1399
 9. Muhammad J, Sinnott RO (2009) Policy-driven patch management for distributed environments. In: 2009 third international conference on network and system security. IEEE, pp 158–163
 10. Chang CW, Tsai DR, Tsai JM (2005) A cross-site patch management model and architecture design for large scale heterogeneous environment. In: Proceedings 39th annual 2005 international carnahan conference on security technology. IEEE, pp 41–46
 11. Lee SW, Kim YJ, Sohn TS, Moon JS, Seo JT, Lee EY, Lee DH (2004) Design the normalized secure patch distribution & management system. *Korean Inst Inf Sci Eng* 31(2):502–504
 12. Midtrapanon S, Wills G (2019) Linux patch management: with security assessment features. In: 4th international conference on internet of things, big data and security, pp 270–277
 13. Zheng J, Okamura H, Dohi T (2019) Security evaluation of a VM-based intrusion-tolerant system with pull-type patch management. In: 2019 IEEE 19th international symposium on high assurance systems engineering. IEEE, pp 156–163
 14. Antonopoulos AM (2014) *Mastering Bitcoin: unlocking digital cryptocurrencies*. O'Reilly Media, Inc
 15. Fan X, Chai Q (2018) Roll-DPoS: a randomized delegated proof of stake scheme for scalable blockchain-based internet of things systems. In: *MobiQuitous '18 proceedings of the 15th EAI international conference on mobile and ubiquitous systems: computing, networking and services*, pp 482–484

A Suggestion for ERP Software Customization Model Using Module Modification Factors



Byung-Keun Yoo and Seung-Hee Kim

Abstract Although many companies have implemented ERP systems to standardize and advance their business practices, in actual operation of such systems companies have been facing challenges such as decreasing numbers in informatization index and lagging productivity. The current study analyzes cases of companies that have implemented ERP to understand customization factors required for successful introduction and operation of the system. By utilizing data analysis technique to understand the features of the deduced ERP customization factors, the current study develops an evaluation model that leads to a model for customizing ERP software modules. The current study will not only provide practical implications regarding the customization of ERP systems but also provide a theoretical basis that supports rational decision-making in the customization process.

Keywords ERP · Package software · Software customization

1 Introduction

Many companies in Korea and abroad have implemented and are operating Enterprise Resource Planning (ERP) systems [1] to quickly respond to the rapidly changing business environment and enhance their international competitiveness. To meet these objectives, companies that implement ERP systems focus on integrating their corporate and organizational resources and raising their work efficiency by way of analyzing and applying standardized advanced business practices based on ERP processes and methodologies. Successful achievement of business informatization through ERP requires companies to carefully customize [2] the ERP package to

B.-K. Yoo · S.-H. Kim (✉)

Department of IT Convergence Software Engineering, Korea University of Technology & Education (KOREATECH), 1600, Chungjeol-ro, Byeongcheon-myeon, dongnam-gu, Cheonan, Chungnam 31253, Republic of Korea
e-mail: sh.kim@koreatech.ac.kr

B.-K. Yoo

e-mail: bkyoo@koreatech.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_79

563

account for their management strategies, processes, and business characteristics. Customization involves a series of activities [1] in which pre-developed ERP software is adjusted and modified to meet the client company's organizational information system, business management environment, and production environment, thereby satisfying the needs of the company. In particular, companies that implement an ERP system for the first time find customization essential to optimize the software to their corporate work processes. As such, to ensure that the ERP system implementation and operation will succeed, administrators and developers must prevent the existing core success factors from being diluted due to the indiscriminate application of advanced work processes unique to each ERP system. Furthermore, they must efficiently consider workplace requirements and apply them to the standardized format of the ERP package.

Data collected from companies that have adopted ERP, however, indicate that the results have been less than satisfactory: Such companies had lower Information Speed Index (ISI) figures compared to other companies within their parent groups [4]. Many companies that have implemented ERP—as opposed to other traditional systems—have failed in the endeavor because the implementation scope had been shrunk, budget limits were exceeded, deliveries were delayed, and/or education and training were insufficient [5]. In particular, businesses were exposed to risky situations when they failed in the ERP implementation because of the falling productivity and increasing maintenance costs that were added on top of the massive resources invested to launch the system.

The current study investigated companies that have implemented an ERP system to understand the customization module factors required to successfully implement and operate an ERP package. Furthermore, it developed a module customization model for ERP software from these customization factors.

The results of the current study can not only be used to evaluate a customized module's level of impact but also provide theoretical criteria that can support rational decision-making regarding various customization conditions. Furthermore, the current study will provide a new direction to research in the field that has been only dealing with how to minimize customization [6]: The new direction will suggest focusing on the impact of individual ERP customization modules.

2 Preliminary

2.1 Background Research

The research on ERP systems to date includes studies on ERP package selection, factors for the successful operation of ERP packages, ways to implement and expand ERP systems, effects of implementing ERP systems and measurement of such effects, relationship between ERP systems and business performance, and other topics about ERP.

Of note, Kim and Oh [1] suggested a strategy for successfully customizing ERP packages. The study provided the following customization strategies: (1) Analysis of different between ERP and independently-developed system in each phase; (2) customization of various technical standards, procedures, and rules; (3) sharing of customization know-how and cooperation among project members; (4) determination of customization level and classification of customization targets; (5) thorough configuration and establishment of follow-on management measures; (6) operation of an organization that reviews, evaluates, and determines the customization processes; (7) fortifying the preparation phase of the ERP system implementation project; (8) minimization of modifications to the ERP package for re-engineering effect; (9) checking whether the customization designs are aligned with the organizational objectives; (10) utilization of upgrades provided by the ERP package developer to maintain the latest system; (11) establishment of a customization policy and training project members; (12) encouraging user participation in the entire customization process; (13) facilitating accurate understanding of the ERP package features and performance; (14) verification of the customization team members' technical competency.

Jin et al. [2] is a study about factors required for successful implementation of ERP system; it proposed that companies crucially consider project costs, corporate competitiveness, and risk acceptance level when they customize an ERP package. Furthermore, the researchers suggested that the companies develop a risk management plan by reviewing their risk acceptance level and establish a risk management organization to conduct regular risk assessment in each phase of the project. Lee et al. [5] presented the fifteen factors required to successfully execute the ERP system implementation phase and calculated the importance of each factor. A key factor for success was selecting an appropriate ERP package (0.140), which was followed by process-oriented approach (0.115), technical adequacy (0.096), minimization of customization (0.083), integration and interfacing with partner company systems (0.076), interfacing with legacy systems (0.068), and support from and participation by the executive management (0.063), in descending order of importance.

Ha and Ahn [7] researched evaluation indicators and impact analysis of public corporations and private businesses that customized an ERP package; the study provided evaluation indicators for each phase of ERP implementation up to stabilization and verified their utility through case studies. As indicated above, no researcher has studied key module factors impacting the customization of ERP systems. There is a dearth of research regarding the customization of the ERP system as well. As such, the current study sets out to discover factors of ERP system customization by analyzing private companies and public corporations of similar sizes that have implemented an ERP system. Based on the deduced factors, the current study will develop and present a model for customizing ERP software.

3 ERP Software Customization Model Using Module Modification Factors

3.1 Overview

The current study looked at private businesses and public corporations that implemented the same ERP system and customized it. Key module factors that were customized for operation, environment, and stability were analyzed by classifying, clustering, and associating them. Based on the analysis results, a fact data comparison was conducted to predict the stability of the systems, which generated key modification factor parameters, which were used to build a model for ERP software customization, in turn.

3.2 Research Procedure

To this end, four private companies and three public corporations that customized their ERP system upon implementation were identified. They were evaluated in terms of ERP system implementation and operation. Private companies were further categorized in terms of their industry and business, into chemical companies (Yuhan Chemical and Wooree Bio) and electronics parts manufacturing (Sungho Electronics and Bluecom). All of the public corporations in the study (Korea Institute for International Economic Policy, Korea Forestry Promotion Institute, and Korea Institute of Patent Information) were supervised by different government agencies but had similar scope of ERP system implementation (all three corporations linked their ERP system to their groupware programs). Private companies in the manufacturing industry were studied in consideration of the fact that their previous systems were upgraded to the current ERP systems; the current levels of stabilization in their systems were also analyzed.

Step 1: Classification of Customization Cases

Cases of companies that customized their ERP systems upon implementation were surveyed; sample companies were extracted based on their similarities and then classified into private companies and public corporations before the data about their customization activities were collected and refined.

Step 2: Evaluation of ERP Operation

The operation of the implemented ERP systems was evaluated using various indicators. In this process, private companies were classified based on their industries and similarities of business processes, into chemical companies and electronics parts manufacturers. Public corporations were analyzed based on the similarities of how their ERP systems were interfaced with their respective existing systems, and then similar cases of ERP system implementation were selected for the study.

Step 3: Analysis of ERP System Stabilization

Upgrade environment and current stabilization of the ERP systems were analyzed. To measure the degree of stabilization, the systems were evaluated in terms of ISO/IEC25000 (software product quality requirements and evaluation) and quality of use. More specifically, the maintainability quality of the ERP software was assessed through detailed indicators such as modularity, reusability, analyticity, modifiability, and testability. Furthermore, satisfaction indicators for trust, satisfaction, and utility were used to evaluate the quality of use of the software.

Step 4: Determination of Software Customization and Modification Factors

Information systems with superior stabilization scores (product quality and quality of use) that were discovered in Step 3 were mapped onto the operability evaluation indicators developed in Step 2, with a goal of deducing the customization features of each module.

In this process, the deduced module modification values were analyzed for clustering, classification, and association; the analyzed data are used to execute the stability prediction data comparison.

Step 5: Proposal of an ERP Software Customization Model Using the Modification Factors

The characteristics of modification factors for the customization modules, as verified in Step 4, were used as parameters in the newly developed ERP software customization model.

4 Conclusion

The current study investigated Korean companies that implemented the ERP system, specifically collecting and refining data about the customization activities of the implemented ERP systems. Next, the operation of the ERP systems was evaluated; afterward, the resulting data were mapped onto the indicators of system stability to develop and propose an ERP software customization model utilizing the software modification factors. The current study is expected to provide a theoretical basis that can support rational decision-making for the software customization process.

References

1. Kim BG, Oh JI (2000) A strategy on the successful customization of ERP packages. *Asia Pac J Inf Syst* 10(3):121–143
2. Jin CH, Kwon Y, Jik, Cui J, Lee SH, Kim SY (2011) An empirical study of ERP systems customizing and performance of SMEs in Korea and China. *J Korea Ind Inf Syst Res* 16(5):127–139

3. Appleton E (1997) How to survive ERP. *Datamation* 3:50–53
4. Lee YM (2011) A study on the problems and handling measure of ERP usage of Korea small and medium companies—The cases of small and medium business. *Korea Logist Rev* 21(3):179–199
5. Lee SG, Kim JJ (2016) An analysis of the importance of the success factors in implementation stage of ERP system. *J Korea Soc Comput Inf* 21(12):165–171
6. Son SH, Ha SG, Kim SS, Kim SW (2011) Analysis of research trends on domestic ERP. *Indus Econ Res* 24(4):2323–2341
7. Ha YM, Ahn HJ (2017) Development of indices for stability of ERP systems in the post-implementation stage, and a case study. *J Korean Inst Inf Technol* 15(1):11–23

A Location-Based Solution for Social Network Service and Android Marketing Using Augmented Reality



Jun-Ho Huh and Yeong-Seok Seo

Abstract This study aims to make it possible for one to share different feelings at the same location through a location-based SNS which is dissimilar to the existing SNS systems. A common platform such as a neighborhood store explorer guides the user by marking the nearby stores based on his/her location but the proposed application can provide some additional information including the reputation or relevant facts through augmented reality (AR) when he/she photograph a specific mark or image with a mobile device for recognition. A location-based Android marketing solution utilizing Vumark AR is proposed in this study where a similar operation method used for the existing RFID and QR code-based systems is adopted to the Vumark-based approach. Vuforia API is used as the main library to store the target marks or images into the database for recognition and present necessary information on a mobile device screen when required.

Keywords Solution · Social network service · Android · Augmented reality · Vumark · Vuforia

1 Introduction

Currently, Social Network Services (SNSs) have become personal media or community where individuals can express and share their individuality, feelings or ideas, not just a means of online socialization or acquaintance [1]. Following the rapid progress of informatization, a variety of SNSs are being continuously introduced in the market but the users are beginning to feel the tedium of their similar basic systems and for this reason, a new type of SNS system has been developed, the aiming to allow the

J.-H. Huh

Department of Data Informatics, Korea Maritime and Ocean University, Busan, Republic of Korea
e-mail: 72networks@kmou.ac.kr

Y.-S. Seo (✉)

Department of Computer Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
e-mail: ysseo@yu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_80

569

users to share different types of feelings, emotions, or senses in the same location or environment. Instead of using a method often used by the existing SNSs, which merely uploading images or indication the locations, this new system employs a new technology—a location-based SNS using Vumark AR as an Android marketing solution.

2 Related Research

Vuforia API is one of the major libraries for such use where the images or a Vumarks to be recognized by the application will be stored in a database by using an imaging tool [2, 3]. This system has been designed in a way to let the users have more fun compared to the others by allowing them to acquire information by shifting their locations and saving it in a more interesting form, instead of just simply searching the information of a certain store and saving it to their mobile devices [4–6].

Augmented reality (AR) is a technology which provides real [7]—world data to the users after combining it with virtual data. Vuforia API is being used in this study as an SDK as it provides convenience in recognizing or processing images. Also, the Vumarks available in Vuforia allow users to experience an AR.

3 Location-Based Solution for Social Network Service

The descriptions of each element of Vumark are as follows: Contour is a part the Vuforia imaging algorithm detects first to find the identifier. Border defines the hexagonal form of a Vumark. Clear Space plays the role of distinguishing the sections (outer and inner) at the boundary while assisting the algorithm in detecting the contour. Code Elements are used as an identifier which contains the information about the data length or type and the number of the elements. Finally, Background can be regarded as a design element of Vumark and has no influence on target detecting.

In an early stage, SNSs adopted a method where their participants communicate with each other after forming a group or an organization sharing similar hobbies or interests. Recently, however, the number of SNS users is increasing rapidly following the increased desire for self-expression. This means that the users were originally using SNSs for socializing purpose but now they are using them as a tool for sharing their individual interest or individuality. Currently, there are closed-type SNSs in which only friends can participate and open-type SNSs all the people can see one's own posts through the function called 'Hashtag'. The social impact of new SNSs is quite large as the posts of others can be included and exposed in one's feed through the 'Share' function. However, the SNS system we are pursuing adopts the existing SNS system but attempts to differentiate it with others by exposing the locations and images (photos) only.



Fig. 1 The web execution screen

The Web page was organized by using HTML and the following map API was used as a map API for the development: a function which is able to output a picture at the center of the screen and shifting and attaching a picture at a specific location at the marker for its registration was created. Also, to add the meaning of ‘attaching a post’ to the map, a ‘post-it’ image (i.e., removable post) was inserted in the background. Figure 1 shows the web execution screen.

We have implemented an SNS by using a location-based service and given the name ‘Whispers’ as people or close friends often talk in a whisper when there are things to talk about intimately. This SNS differentiated itself from the others by uploading pictures and GPS-based location information only. Big data can be generated for at the same spot through this SNS as the location data includes latitude & longitude in the database when uploading pictures.

Meanwhile, the Google map API was used when starting the development process but it was found later that the Daum map was more efficient and accurate when setting a location as well as finding the names of buildings or areas. For this reason, the latter was selected to increase accuracy. Also, we have tried to make the SNS more readable by showing only the necessary information in the webpage. The below picture is showing its execution screen.

The service proposed in this study is a location-based Android marketing solution using Vumark AR adopting a similar operation methodology often used for the RFID or QR code recognition. Vuforia API was used as the main library to store a particular target mark or image in the database for the recognition by the application. This method allows the system to provide necessary information to the users on their mobile devices in an augmented form that has not been available.

As mentioned earlier, the user lets his/her mobile device to recognize the Vumark attached to the store with its camera module and output the store information or others on the device based on the recognized image or mark (Fig. 2).



Fig. 2 Implementation of Vumark AR

One DB is used for the recognition of the mark and another is used to store the store information and construct the review board data. Different from a static object, one of the objects to be outputted takes the shape of a virtual web browser and the user can leave a review.

In Fig. 2, the symbol (mark) of our university was set as a target image and let us suppose that the above picture is the mobile screen of the user. In this case, an additional event-processing can be performed to recognize the user input in a virtual environment. Also, by entering a code into the object to be outputted, a movement can be designated to it or use it as a button instead.

4 Conclusion and Future Work

Research on SNSs was discussed in this study based on a new system along with the system development. In our modern society, these SNSs have become a window on which individual personalities, emotions, or sensibilities are reflected or shared as a personal media or community, not just a place for simple socialization. Development of a new type of SNS was attempted by using the space for personal expression as a location for sharing these feelings. Prior to implementing the service, a survey questionnaire was distributed to 150 people to check and analyze their requirements. Most of the respondents had a positive response toward our method and many of them were relatively feeling tedious toward existing SNSs.

The service proposed in this study is a location-based Android marketing solution using Vumark AR adopting a similar operation methodology often used for the RFID or QR code recognition. Vuforia API was used as the main library to store a particular target mark or image in the database for the recognition by the application. This method allows the system to provide necessary information to the users on their mobile devices in an augmented form that has not been available.

Acknowledgements This work was supported by the This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00743) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). Also, this work was supported by the National Research Foundation of Korea (NRF, grant no. 2017R1C1B5077157) funded by the Korean government (MSIT).

References

1. Rau P-L, Gao Q, Ding Y (2008) Relationship between the level of intimacy and lurking in online social network services. *Comput Human Behav, Elsevier* 24(6):2757–2770
2. Ibañez AS, Figueras JP (2013) Vuforia v1. 5 SDK: analysis and evaluation of capabilities. Master in Science in Telecommunication Engineering & Management
3. Peng F, Zhai J (2017) A mobile augmented reality system for exhibition hall based on Vuforia. In: 2017 2nd international conference on image, vision and computing (ICIVC). IEEE, pp 1049–1052
4. Adrianto D, Hidajat M, Yesmaya V (2016). Augmented reality using Vuforia for marketing residence. In: 2016 1st international conference on game, game art, and gamification (ICGGAG). IEEE, pp 1–5
5. Purnomo FA, Santosa PI, Hartanto R, Pratisto EH, Purbayu A (2018) Implementation of augmented reality technology in sangiran museum with vuforia. In: IOP conference series: materials science and engineering, vol 333, No 1. IOP Publishing, p 012103
6. Kim T-J, Huh J-H, Kim J-M (2018) Bi-directional education contents using VR equipments and augmented reality. *Multimedia Tools Appl* 77(22):30089–30104
7. Feiner S, Macintyre B, Seligmann D (1993) Knowledge-based augmented reality. *Commun ACM* 36(7):53–62

Artificial Intelligence Based Electronic Healthcare Solution



Seong-Kyu Kim and Jun-Ho Huh

Abstract One of the major keywords in the current digital world is Artificial Intelligence (AI) which is playing a major part in all kinds of advanced service systems, offering more convenience, better efficiency/effectiveness by controlling system hardware intelligently in a way humans never experienced. AI is also playing an essential part in the healthcare or bioelectronics industry where enhanced service function and sophistication have become a critical factor in a keen completion. Thus, this paper focuses on its contributing factors to human society and provides an opportunity for the discussions on the relevant convergence technologies.

Keywords Artificial intelligence · Healthcare · Electronic healthcare · Solution

1 Introduction

With the breakthrough of data and computation science, big data related methodologies such as blockchain, deep learning, AI have proven to be powerful in discovering some hidden information underlining natural events, which can't be directly observed by other methods. Big data analytics greatly facilitate people's exploration of science knowledge [1]. The most distinctive achievement of using advanced data science technologies is the AlphaGo, the computer program integrating AI trained by deep learning method. It has been defeating the top Go game players, who had been thought to be invincible against computer program before the born of AlphaGo [1, 2]. Figure 1 shows Google DeepMind Challenge Match. This achievement shed

S.-K. Kim

Department of Information Security, Joongbu University, Gyeonggi-do 10279, Republic of Korea
e-mail: guitar77@gmail.com

Department of IT Strategy, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea

J.-H. Huh (✉)

Department of Data Informatics, Korea Maritime and Ocean University, Busan, Republic of Korea
e-mail: 72networks@kmou.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_81

575



Fig. 1 Google DeepMind challenge match

lights on the emerging field to computing system researchers globally, who seek to build high fidelity models of computing systems.

2 Artificial Intelligence and Big Data

Artificial intelligence refers to a next-generation system for more efficient handling of difficult human processes using machine learning and deep learning. Artificial intelligence is a branch of computer science that studies how intelligent behavior in humans can be applied to mechanical systems.

In addition, artificial intelligence makes learning data intelligent by continuing learning. These tasks are called tuning tests. Turing tests are tests that determine whether a machine can operate intelligently as a person. The test was designed in 1950 by Alan Turing, the father of modern computer science.

This test requires a human evaluator to interview machines and people without hearing and visual contact. Both the machine and the person interviewed claim to be human. If the assessor fails to distinguish people from the machine, the machine passes the tuning test and the machine is considered to have the same intelligence as humans. And this is not the only thing that can be learned from the history of artificial intelligence. Looking at the evolution of artificial intelligence, there is no such thing as all-inclusive artificial intelligence. Rather, artificial intelligence is a collection of techniques and techniques that are closely related and connected to each other. Figure 2 shows a concept image of what the banknote could look like.

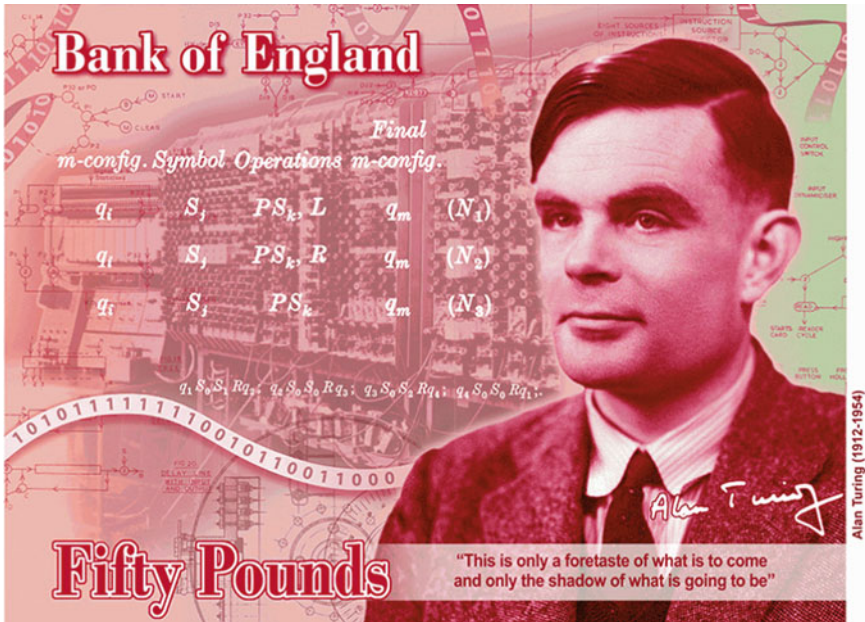


Fig. 2 A concept image of what the banknote could look like (Bank of England)

The two most well-known technologies are Machine Learning and Deep Learning, while Deep Learning is a sub-field of Machine Learning. In machine learning, algorithms acquire knowledge based on “experience,” or sample data. The system identifies and integrates critical features to establish its own rules to apply to unknown data. One example of machine learning is the recommended function of the video portal. As the user’s responses to video recommendations are analyzed, the recommendation function is improving day by day. And deep learning is based on the analysis of large amounts of big data. The system classifies and identifies digital information from various sources. Although the learning process does not use all data at the same time, it continues to use a new collection of information. The goal is to classify and grasp what has been learned. As such, artificial intelligence talks about the process of continuously learning the data of Big Data.

A large 1st dimensional dataset is referred to as big data and such a dataset is often generated from the case where the constrains on an available data form has been loosened, including the structured data such as taking an HTML, XML or weblog form; the semi-structured data forming a metadata or schema; the structured data, typical form of which is video data [2–5]. Also, big data technology is emerging as an extension of technologies or resources dealing with these types of data. Data has been largely extended due to the increase in data volume and diversification so that a

database technology which is able to process such a data faster than the existing ones became necessary. Meanwhile, Prof. Y. Pan (Georgia State University) defined the big data technology in his keynote speech at WITC 2019 [6–8] that it is a technology which allows us to see a big picture by putting the pieces of the puzzle together. That is, big data technology is a technology which extracts a certain value from a large dataset (dozens of terabytes) including not only structured but also unstructured data for analysis [9].

3 Blockchain

Blockchain refers to a decentralized architecture that deviates from a central system. In other words, blockchain is a data distribution processing technology. This refers to the technology that all users who participate in a network store data, such as all transactions, in a distributed manner. This stored data is called blocks, and blocks are grouped in order of time [10].

All of these users have transaction details, so when checking the transaction details, all users must check the books they have. For this reason, blockchain is also called public or distributed transaction books. The existing transaction method stores all transactions at the central bank, ie the bank.

This is because we have to prove the transaction between individuals by storing. Unlike banks, blockchain will be stored by people who participate in the network. If there are 1,000 people involved in a network, create 1,000 blocks of transactions between individuals, send them to all 1,000 people, and save them. Later, when checking the transaction details, the stored data are verified by dividing them into blocks. As mentioned above, blockchain is characterized by distributed storage. Under the existing transaction method, the central server is attacked in order to falsify the data.

However, blockchain has multiple people storing the same data, making it difficult to do so. It is considered virtually impossible to hack a blockchain network because it requires attacks on all participants' transaction data in order to tamper with it. As such, blockchain is designed based on decentralization and reliability, so blockchain does not require a central administrator. Decentralization is possible because many can store and prove data without central agencies or managers. This concept is also called blockchain.

The blockchain is also called a public transaction book. It is a technology that prevents double payment that can occur in financial transactions and can not be tampered with. It has become a core technology of PINTECH in Korea [11].

4 Artificial Intelligence Based Electronic Healthcare Solution: Personal Health Record dApp

The Personal Health Record (PHR) is the individual health record owned by hospitals [11]. Most hospitals around the world now use the Electrical Medical Records (EMR) system. By digitalizing hospitals’ medical records and providing a PHR for the blockchain service, we can support each individual in recording, managing, sharing and controlling them, and pursue medically required records in compliance with the national interoperability standard. The government has established the national medical system at three levels—1st (small hospitals), 2nd (mid-sized hospitals) and 3rd (large hospitals). It is only through this procedure that users can receive the benefits of medical insurance. When sending a patient’s information from a class 1 hospital to a 2nd or 3rd class hospital, the blockchain can be used to apply original note checking and data encryption, allowing data to be used in perfect safety and security. In particular, the Autochain Blockchain’s medical record is an XML document that complies with the W3C standards for HL7 medical services and uses machine-learning metadata, thus allowing personal data to be recycled easily in the future. Figure 3 shows personal health record model flow.

In addition, by using Autocombine XML technology, the blockchain service can be received with a UI/UX (User Interface/User Experience) view on the browser of a specific user’s device by viewing the PHR (Personal Health Record) document. Figure 4 shows personal health record model UI/UX.

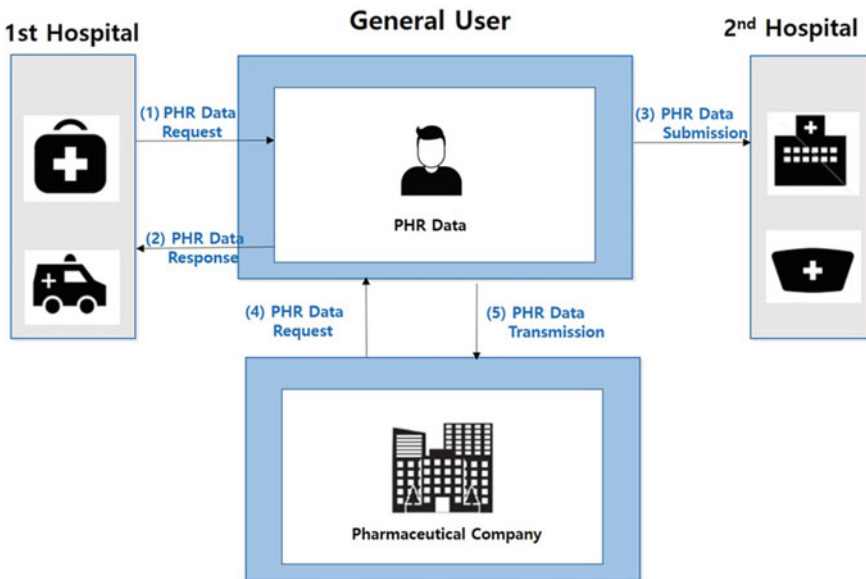


Fig. 3 Personal health record model flow

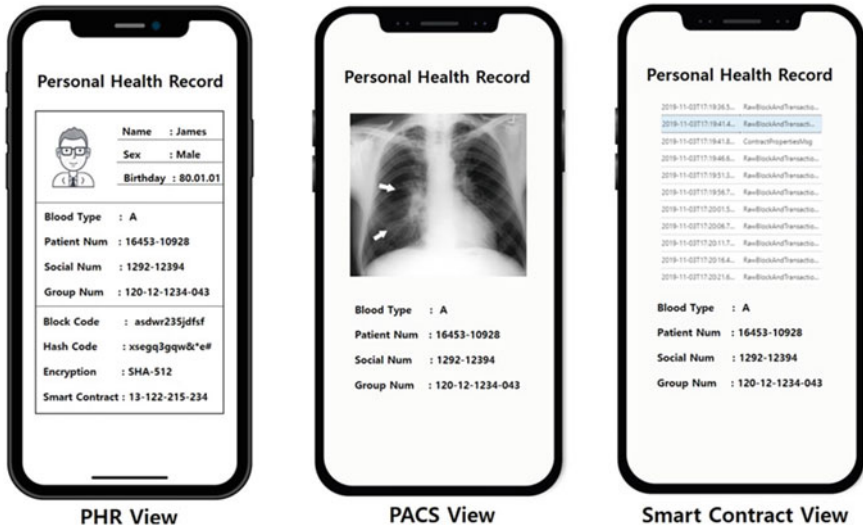


Fig. 4 Personal health record model UI

5 Conclusion

One of the major keywords in the current digital world is AI, which is playing a major part in all kinds of advanced service systems, offering more convenience, better efficiency/effectiveness by controlling system hardware intelligently in a way humans have never experienced. Thus, the AI blockchain has great potential to change traditional industries through key characteristics such as decentralization, data illegality, anonymity, traceability, and so on. In this paper, we propose a block chain based on an automation engine that can easily create and apply a block chain to an automation block chain model called Autochain by complementing the problems of the block chain structure and the blockchain.

Acknowledgements This work was supported by the This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00743) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). Also, this work was supported by the National Research Foundation of Korea (NRF, grant no. 2017R1C1B5077157) funded by the Korean government (MSIT).

References

1. Bhattarai BP, Paudyal S, Luo Y, Mohanpurkar M, Cheung K, Tonkoski R, Manic M (2019) Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions.

- IET Smart Grid, IET 2(2):141–154
2. Porter J, Alan turing is the face of UK's new £50 note. <https://www.theverge.com/2019/7/15/20694453/uk-50-banknote-alan-turing-announcement-scientist-computer>
 3. Wu X et al (2013) Data mining with big data. IEEE Trans Know Data Eng, IEEE 26(1):97–107
 4. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mobile Netw Appl, Springer 19(2):171–209
 5. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst, Springer 2(1)
 6. Zeng M, Li M, Fei Z, Wu F, Li Y, Pan Y, Wang J (2019) A deep learning framework for identifying essential proteins by integrating multiple types of biological information. IEEE/ACM Trans Computat Biol Bioinf
 7. Ruan C et al (2019) PTCP: a priority-based transport control protocol for timeout mitigation in commodity data center. Future Generat Comput Syst, Elsevier
 8. Pan Y, WITC 2019. <https://www.worlditcongress.org/2019/>
 9. Park C-K, Huh J-H (2019) What is big data? A method of collecting or analyzing big data In: Fall proceedings of the Korean data analysis society
 10. Kim SK et al (2018) A study on application method for automation solution using blockchain dApp platform. In: International conference on parallel and distributed computing: applications and technologies. Springer, Singapore, pp 444–458
 11. Kim S-K, Huh J-H (2019) A study on mainchain and sidechain for blockchain development automation solution, MITA

Optimal Location Recommendation System for Offshore Floating Wind Power Plant Using Big Data Analysis



Sang-Hyang Lee and Jun-Ho Huh

Abstract The necessity of offshore wind power plants has emerged due to the issues involving low-frequency sound or noise-related health problems, socio-economic or environmental problems. It is not easy to determine the optimal location for the wind power plants: First, as we can see from the overseas cases, the safety must be considered first followed by economic and environmental problems. Since these plants should not become an obstacle for traffic emergencies, they should not be constructed in the vicinity of a civil or military airport and at the same time, the existence of any nearby fish farms should be checked as well as the plants might largely affect their economy. Thus, this study attempts to provide the most efficient and preferable service by identifying and informatizing the issues related to the changes in the navigation routes of vessels or the condition of individual fish farms based on the big data accumulated over 30 years and realistic simulations. In conclusion, this study aims to find the optimal location for an offshore floating wind power plant.

Keywords Wind power plant · Big data · Big data analysis · Location · Ocean

1 Introduction

Wind power generation transforms the energy generated by the wind turbine blades into electric energy through the generator and its system can be largely divided into two types: a pinwheel-shaped horizontal-axis wind power system we often imagine or a-vertical axis wind power system which is often installed vertically to the ground [1, 2]. Wind energy was used in the seventh-century using the windwheel developed by Heron (Alexandria) who made the blades with reeds or cloth and used it for many purposes including pumping up water, milling, and sugar manufacturing. Then, in

S.-H. Lee · J.-H. Huh (✉)

Department of Data Informatics, Korea Maritime and Ocean University, Busan, Republic of Korea
e-mail: 72networks@kmou.ac.kr

S.-H. Lee

e-mail: huri2017@naver.com

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 715,
https://doi.org/10.1007/978-981-15-9343-7_82

583

the nineteenth century, windwheels were used to generate electricity in Denmark to power a mill or pump. The modern wind power generator was originally developed in the US in 1888 [2, 3].

The wind power generation was achieved in the Republic of Korea (ROK) since the early twenty-first century and some of the famous onshore wind farms are Daegwallyeong, Jeju, Taebaek, and Yeongdeok wind farms. It is being considered that the average wind speed should exceed 7 or 7.5 m/s to be regarded as economically feasible but space/site which would satisfy such a requirement is quite scarce in the ROK as about 70% of its entire territory is mountainous. Therefore, since the problems involving low-frequency noise, economic/social/environmental issues have become a major topic in various media, offshore wind power generation has emerged as an alternative.

The wind power generation systems in the ROK are being constructed mainly on a large lake or the shore to generate electricity and can be largely divided into four types: Concrete caisson, monopile, Jacket, and floating. Among them, the floating type is considered to be the most suitable model for the future as it can be constructed on shallow waters and the risk of current-related accidents is relatively lower than the onshore wind power generation as it generates and transmits electric power through its own mooring system or underwater cables.

In general, offshore wind power generators produce lesser noise than the onshore generators but can generate power more efficiently as the offshore winds are often stronger. Currently, a number of large-scale wind farms including Seonamhae, Daejeong (Jeju), Seosaeng (Ulsan) farms are under construction but its process is slowing down due to the protest by the fishermen, civic or environmental groups. For this reason, this study attempts to recommend the most suitable site for an offshore wind power plant through big data analysis considering tidal currents, conditions of fish farms or ecosystem, emergency landing of aircraft, ships' navigation routes, and public opinion.

2 Related Research

The 4th Industrial Revolution has entered a new phase where some of the new technologies involving big data analysis, artificial intelligence (AI), IoT, etc. are being innovated. 'The 4th Industrial Revolution' is the product of the advancement of big data and AI technologies which allow one program to replace hundreds of machines, compared to the past where one machine replaced hundreds of workers [4–6].

In this regard, the characteristics of the first keyword 'Big Data' can be described as 3V: Velocity, processing and analyzing a large volume of data quickly; Variety, data can be classified as a structured, semi-structured, or unstructured data; and Volume, being large-scale data. There are some popular big data analysis techniques such as text mining, clustering, opinion mining, etc. and these are being used widely in the field of politics, economy, sports, etc. The second keyword 'AI' refers to

an intelligence created by the system and has its own unique learning technique: Machine Learning, an autonomous learning process utilizing input data based on the basic rules given initially [7–9]. The artificial neural network modeling the human neuron structure is one of such learning models; and Deep Learning, dealing with the artificial neural network in which artificial neurons are piled up and connected during the input and output processes [10–12].

This paper attempts to recommend the most suitable site for an offshore wind power plant through big data analysis.

3 Optimal Location Recommendation System for Offshore Floating Wind Power Plant

Figure 1 is a proposed App that checks a desirable location on which to build a floating wind power plant. Once it is launched, four menus will appear: With the first menu Ocean Current confirms the wind speed (strength) based on the ocean current; The second menu Fishing Ground visualizes the movement of fish shoals; The third menu Sea Route allows the user to locate a ship by connecting with the site indicating ship’s navigation route; the last menu sets and shows locations of military or civilian airports and their surrounding areas in each region as potential emergency aircraft landing sites. Then, when a region is selected in the menu Recommend Land, the app confirms the suitability of the region considering above four conditions and a function recommending an alternative site will be available as well.

Meanwhile, the UML used for selecting a suitable site is shown in Fig. 2. Also, Fig. 3 shows a system diagram for selecting a suitable site for a floating wind power

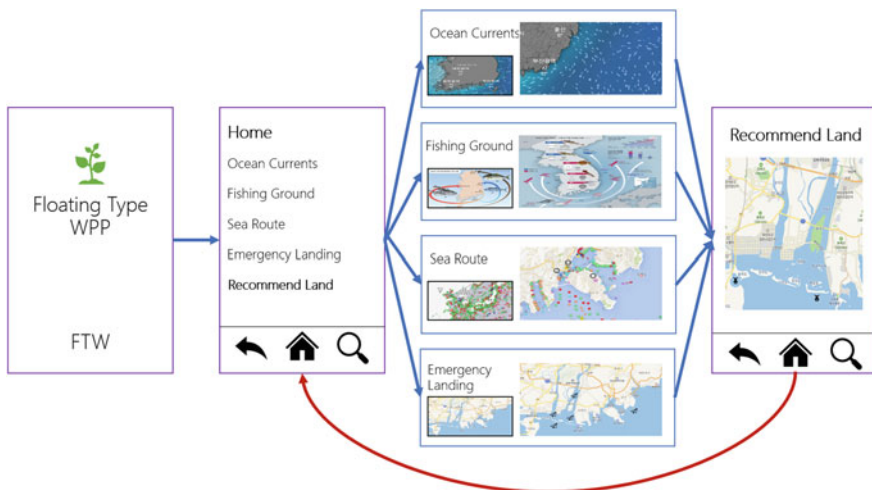


Fig. 1 An app used to select an appropriate site for a floating wind power plant

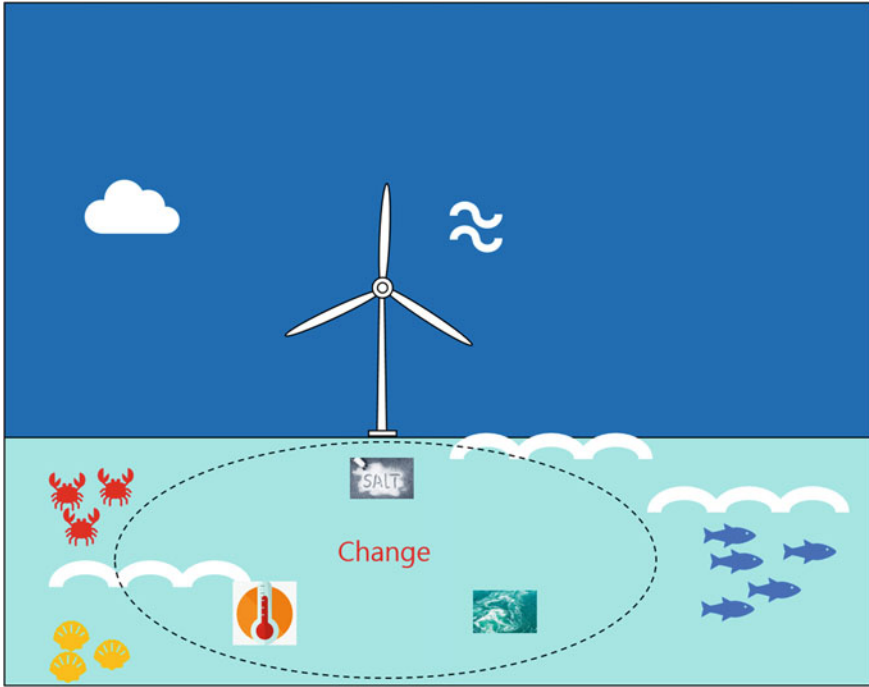


Fig. 2 A UML guide map for selecting a suitable site for a floating wind power plant

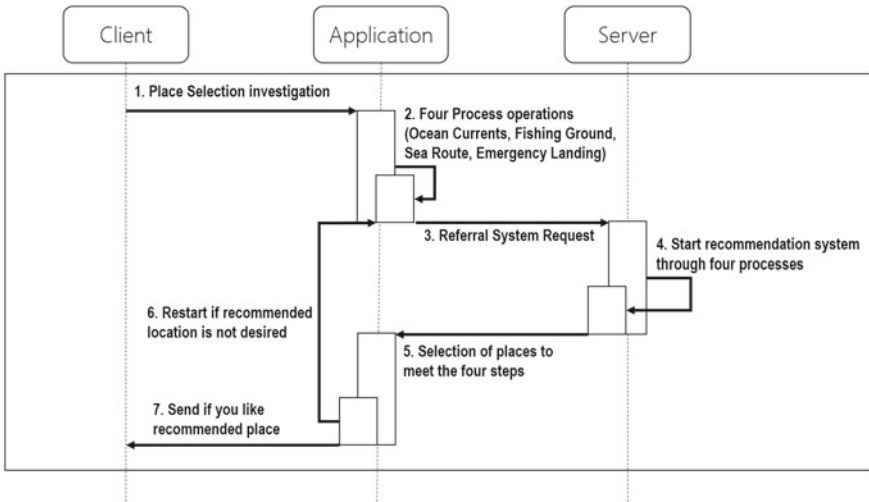


Fig. 3 A system diagram for selecting a suitable site for a floating wind power plant

plant. If the picture of a windmill is considered as a site where the wind power plant will be constructed on, the salinity, current, and temperature within the radius of the circle could change and in some cases, the fishes in the picture would die due to a considerable change in their ecosystem so that it is essential that such an ecosystem should be checked carefully when building a wind power plant.

4 Comparison with Other System of Republic of Korea

4.1 Airport Transfer at ROK

In 2011, there was a nationwide debate over a new airport site which was to replace the existing heavily congested Gimhae Airport in Yeongnam region. The candidate sites were Miryang City (Gyeongnam) and Gadeokdo (Busan). At that time, the candidates in an election made different promises that ended up as a national issue, which led to a rapid increase in the land values of these two regions. The public opinion in the Busan area was promoted in favor of Gadeokdo by revealing the result of a survey conducted in the Yeongnam region, which had shown over 70% of citizens in that region preferred Gadeokdo over Miryang City to which a similar claim was presented by the residents of Gyeongnam and Ulsan areas. During this period, both sites experienced the speculation frenzy but it eventually calmed down as the government decided to expand the Gimhae Airport instead of building a new one in 2016. In the end, the speculators failed and the respective land values decreased to the original prices.

4.2 Ulsan Offshore Power Plants by Wind Energy at ROK

Ulsan City is currently proceeding with an offshore wind power plant construction project in the Seosaeng-myeon and Gandong (Buk-gu) districts after being approved by the government. The offshore wind power plant in Seosaeng-myeon is being constructed with a method of recycling the existing gas production facility so that has no direct and close relationship with the land price. However, this project is being suspended due to the compensation problem with fishermen. Also, the real estate price dropped in the Gandong district after the government announced that the plant will be constructed about 1–2 km from the shoreline.

4.3 Jeju Offshore Power Plants by Wind Energy at ROK

Ten wind turbine generators are being operated on the sea about 600–1,200 m from the Dumo-ri and Geumdeung-ri shorelines (Hankyeong-myeon, Jeju City) and as the operating company is returning a portion of their annual profit to the residents, they seem to be wanting an expansion of that business. On the other hand, Daejeong-eup (Seogwipo City) is planning to go ahead with the Moseulpo Port expansion project as a part of their national ports development project but it is now tumbling down after the public opposition for proceeding with an offshore wind power generation near the port has been announced.

4.4 Southwest Sea Large Scale Offshore Power Plants by Wind Energy at ROK

As a large-capacity remote offshore wind farm having an offshore substation located about 10 km from the Buan-gun shoreline (Jeollabuk-do), the Seonamhae wind farm testbed was supposed to be completed by 2014 but has been delayed for five years. Also, the projects involving the construction of a demonstration site and its expansion have become quite difficult due to strong opposition by fishermen and local community, in addition to continuous negative public opinion.

Meanwhile, Fig. 4 is showing the web crawling result from the keyword search in the Naver News section. The issues related to wind farms were used as keywords and searched through in the news published during the period from April 2018 to November 2019 for analysis creating a word cloud. It was considered that the larger the size of the word, it appeared more often in the news so that from the perspective of ‘public opinion’, which is being focused in this study, such a word should be considered in priority when proposing a wind power generation farm.

The issues involving the relocation of an airport (4.1) always attract people’s attention as the land price might rise rapidly. Also, those who have a piece of land in the region would sometimes try to promote public opinion to support and drive the project. Thus, we’ve attempted to check the facts about the project by web crawling the news articles. In conclusion, the reason for the slow introduction process of a wind power plant was related to the problems associated with fishery resources, land prices, etc., all of which were closely connected to the local residents. Therefore, in order to proceed with the project smoothly without any setbacks, it is essential to create a win–win situation with them or the local community by focusing on the words frequently appearing in the news in advance (Fig. 1).

a wind power plant. This will definitely reduce unwanted conflict with the local community.

Acknowledgements This work was supported by the This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00743) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). Also, this work was supported by the National Research Foundation of Korea (NRF, grant no. 2017R1C1B5077157) funded by the Korean government (MSIT).

References

1. Korea Wind Energy Industry Association (in Korean)
2. Domestic and overseas wind power generation industry and technology development, 2018 Technical report from Korea Wind Energy Association, 2018 (in Korean)
3. 2018 preliminary feasibility report, floating offshore wind power technology empirical project, 2018 (in Korean)
4. Korea offshore wind power web site (West South Offshore Wind Power Development business) (in Korean)
5. New recycling energy data center (in Korean)
6. Tong KC, Technical and economic aspects of a floating offshore wind farm. Elsevier, *J Wind Eng Indus Aerodyn* 74:399–410
7. Fetanat A, Khorasaninejad E, A novel hybrid MCDM approach for offshore wind farm site selection: a case study of Iran. Elsevier, *Ocean Coastal Manage* 109:17–28
8. Laura C-S, Vicente D-C, Life-cycle cost analysis of floating offshore wind farms. *Renew Energy*, Elsevier 66:41–48
9. Huh J-H (2019) Reefer container monitoring system using PLC-based communication technology for maritime edge computing. *J Supercomput*, Springer, USA 1–23
10. Castro-Santos L, Diaz-Casas V (2015) Economic influence of location in floating offshore wind farms. Elsevier, *Ocean Eng* 107:13–22
11. Forinash C, DuPont B (2016) Optimization of floating offshore wind energy systems using an extended pattern search method. In: ASME, 2016 35th international conference on ocean, offshore and arctic engineering. American Society of Mechanical Engineers Digital Collection
12. Lee M-J (2011) Utilization of big data and open data. *Int Inf Secur* 2(2):47–64

Efficient Data Noise-Reduction for Cyber Threat Intelligence System



Seonghyeon Gong and Changhoon Lee

Abstract Preemptive respondents on cyber threats have become an essential part of cybersecurity. Cyber Threat Intelligence (CTI) is an evidence-based threat detection and prevention system. CTI system analyzes and shares the security data to mitigate evolving cyber threats using security-related data. However, to gather enough amount of data for analysis, the CTI system uses various data collection channels. The reliability of data collected from these channels is a critical issue because the inaccurate and vast amount of information could degrade the performance of threat detection. Thus, proper filtering is needed to remove the noise data. In this paper, we propose a data noise-reduction algorithm. The proposed algorithm reflects the contextual characteristics of CTI data and reduces noise data in the CTI dataset. Noise-reduced dataset increases the performance of machine learning and deep learning-based attack prediction models. In our experiment, we conducted a cyber-attack classification using a noise-reduced CTI dataset. As a result, we improve the accuracy of classification from 84 to 96% and reduce the volume of the dataset by 70%.

Keywords Cyber threat intelligence · Noise reduction · Cyber attack · Machine learning

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00158, Development of Cyber Threat Intelligence(CTI) analysis and information sharing technology for national cyber incident response).

S. Gong · C. Lee (✉)

Department of Computer Science and Technology, Seoul National University of Science and Technology, Seoul, Republic of Korea

e-mail: chlee@seoultech.ac.kr

S. Gong

e-mail: gongsh@seoultech.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_83

1 Introduction

As the scope of the network has expanded widely, various devices have been connected to the Internet. This change reads the skyrockets of the complexity of the threats in the cyber environment has also increased significantly. With the concept of the Internet of things (IoT) [1] being applied to various fields, network technology is closely connected with human life, and the attacker's attack point also has been diversified. As a result of these expansions, the damage caused by cyber-attacks is increasing. Therefore, there is a growing demand for technologies that can predict cyber-attacks as well as improve security technologies.

Cyber threat intelligence (CTI) has emerged for preemptive threat prediction and prevention. CTI is a knowledge-based information system aimed at security situation awareness and preemptive response to cyber-attacks [2]. CTI uses a variety of forms of security data to profile cyber threats and provide responses to threats by providing optimal information to security managers. CTI system analyzes the threat-related data such as traffic, log, and threat reports about Indicator of Compromise (IoC) and generates a reputation of IoC. A large number of security-related data is required to enhance the effectiveness and quality of the CTI system. To end this, various data collection channels, such as security information and event management (SIEM), could be used for data collection.

But the increase in the volume of data brings problems about the reliability of data [3]. Data generated by unreliable procedures creates inaccurate threat reports, which can have a significant impact on the performance and accuracy of the system. Therefore, it is essential to verify the reliability of the data.

In this paper, we propose an algorithm, ReputationRank, that removes the noise data through cross-reference analysis between data and prioritize important data. Through the proposed algorithm, the effectiveness of each network resource can be quantified and calculated, and the importance of network resources can be classified qualitatively and quantitatively based on this information. It also reduces the complexity of the dataset by reducing noise data. This algorithm uses the concept of rank between related information [4]. In the experiment, we conducted the data noise-reduction process using a report dataset of 70,885 network resources. As a result, we improve the accuracy of threat detection from 84 to 96%, and we reduced the complexity of the dataset by 70% using our algorithm. In the real field, the amount of data to analyze is a critical point because the complexity of the dataset directly affects the performance and immediacy of the system. Our result shows that the proposed noise-reduction mechanism could be a useful method to enhance the accuracy of detection and availability of the system in the real field.

Section 2 of this paper introduces the studies related to CTI and data reliability. The whole structure of the proposed model is illustrated in Sect. 3. Section 4 presents experiments and results to demonstrate the practical effectiveness of the proposed framework, and we discuss and conclude the meaning of this result in Sect. 5.

2 Related Work

There are many efforts to standardize the CTI system at the organization level. Structured threat information expression (STIX) [5] is a CTI data representation language to share the CTI data efficiently, and represent data in JSON form. Trusted and Automated eXchange of Intelligent Information (TAXII) [6] protocol is a server-client based communication protocol for sharing data expressed in STIX. These technologies are regarded as de-facto standards.

Kim, K., and Kim, H. K. proposed an automated CTI dataset generation system and explained the statistical properties of the generated dataset to alleviate the problems of the CTI system originating from the dataset [7]. Sillaber et al. [8] analyzed the state of technology of various professional organizations from multiple perspectives on data collection, processing, sharing, and storage to examine data sharing technology, a problem of CTI technology.

There was research conducted to verify the reliability of data through cross-validation of the contents of CTI data. In our previous work, we collected CTI data from multiple OSINT channels and evaluated the reliability of the data by cross-validating the contents of the data [9]. Meier's work [10] suggested a way to rank a CTI feed. These researches have evaluated feeds through the content of data provided in CTI feeds and cross-references to other feeds.

3 Proposed Noise Reduction Model

We propose a new noise reduction model for CTI data based on the influence of reputation information. The proposed data noise-reduction algorithm, ReputationRank, reduces unnecessary data in the CTI system using influence evaluation of security data. The proposed algorithm is based on the concept of rank evaluation on related data, the PageRank algorithm [4].

Relations of the security data is expressed as network resources such as IP and domain resolution and distribution history of malware or ransomware from specific network addresses. Thus, to delicately illustrate the relations among security data, time-series characteristics must be included in prioritization. Equation (1) is ReputationRank algorithm for prioritization of each data. We adjust the time-series characteristics by adding timestamp term into the formula in Eq. (2). The closer the current data is to the last data, this term has a higher value to reflect its impact. The constant value of the time-series term adjusts the loss of rank value by moving the expectation of time-series term. The overall algorithm based on the above equations is illustrated in Algorithm 1. In equations, D_i means the i -th resource of dataset D , and $RR(D_i)$ shows the reputation rank value of D_i . d is a damping factor and means the ratio between the volume of dataset and relations in the dataset. $L(D)$ is the size of dataset D , and $IB_j(D_i)$ and $OB_j(D_i)$ means the number of inbound and out-bound relations of

data j to i in dataset D respectively. $FT_j(D_i)$, $LT_j(D_i)$ and $CT_j(D_i)$ means the times-tamp of first-seen, last-seen and reported time of resource i . As the hyper-parameter, the filtering threshold of rank value is adjusted by the repetitions considering the processing time and volume of overall data. Information that has lower rank values than this filtering threshold is removed from the dataset. And we adjust the damping factor value to reflect the average number of relations that each threat data we have.

$$RR(S_i) = \frac{1-d}{L(D)} + d \times \sum_{j=1}^{L(IB(D_i))} T(IB_j(D_j)) \times \frac{RR(IB_j(D_j))}{L(OB(IB_j(D_j)))} \quad (1)$$

$$T(D_i) = \frac{1}{2} + \frac{CT(D_i) - FT(D_i)}{LT(D_i) - FT(D_i)} \quad (2)$$

Algorithm 1: Reputation Rank

Data: D : list of network resource
 d : damping factor
 f : filtering threshold
 t : threshold for recursive process

Result: $RR(D)$: rank value list of D

```

1 for each resource  $D_i$  in  $D$  do
2    $RR(D_i) \leftarrow 1 / L(D)$ 
3 end
4 while  $\delta \leq t$  do
5   for each resource  $D_i$  in  $D$  do
6      $e \leftarrow \emptyset$ 
7     for each item  $j$  in  $IB(D_i)$  do
8        $e \leftarrow e + RR(j) / L(OB(j))$ 
9     end
10     $tmp \leftarrow (1-d) / L(D) + e$ 
11     $\delta \leftarrow \delta + |tmp - RR(D_i)|$ 
12     $RR(D_i) \leftarrow tmp$ 
13  end
14   $\delta \leftarrow \delta / N$ 
15 end
16 for each resource  $D_i$  in  $D$  do
17   if  $RR(D_i) < f$  then
18     remove  $D_i$  from  $D$ 
19   end
20 end
21 return  $RR(D)$ 

```

4 Experiments and Result

To evaluate the proposed noise reduction algorithm, we conduct an experiment using 70,885 IP-related CTI data. These reports had collected from external CTI feeds such as ThreatCrowd, Open Threat eXchange (OTX), and VirusTotal. By preprocessing, we extracted 43,892 relations extracted from the CTI dataset. The dataset is composed of eight attack types, and Table 1 shows the distribution of each attack type. The number of relations divided by the number of data, 0.6192, was used as the damping factor. Also, the first initialized rank value of each data, 0.00000014, is used as the threshold value (1% of initial rank value). After conducting the noise reduction process, we filtered 70% of the overall data.

We performed this experiment under Linux Ubuntu 18.04 operating system, Intel i7-9700 k CPU, 16 GB RAM size, and nVidia GeForce 1080ti GPU environment.

To compare the result between the original and noise-reduced dataset, we had chosen seven classification algorithms for the experiment: Naive Bayes, Logistic Regression, Linear SVM, Neural Network, Nearest Neighbor, Decision Tree, and Random Forest.

Also, 10-fold cross-validation had used to train the dataset [11]. The Fig. 1 shows the result of the cyber threat type detection result. With a noisy dataset, the attack detection ratio of each classification algorithm has quite a low accuracy (under 90%). However, with the noise-reduced dataset, we achieved a 10% improvement in the detection ratio of attack types with all algorithms. Notably, we improved 12% of detection ration with random forest, and it reached 96% of detection ratio.

Table 1 The number of data per attack types

Attack type	Original	Noise-reduced	Reduction ratio (%)
Malware distribution	3,443	1,015	70.52
Pharming	23,853	7,159	69.99
Malware information Leaking	6,275	1,924	69.34
Phishing	27,621	8,151	70.49
Command and control	2,115	649	70.52
Information leaking	1,865	544	69.31
Malicious code distribution	5,536	1,661	70.00
Unknown	177	38	78.53
Total	70,885	21,141	70.18

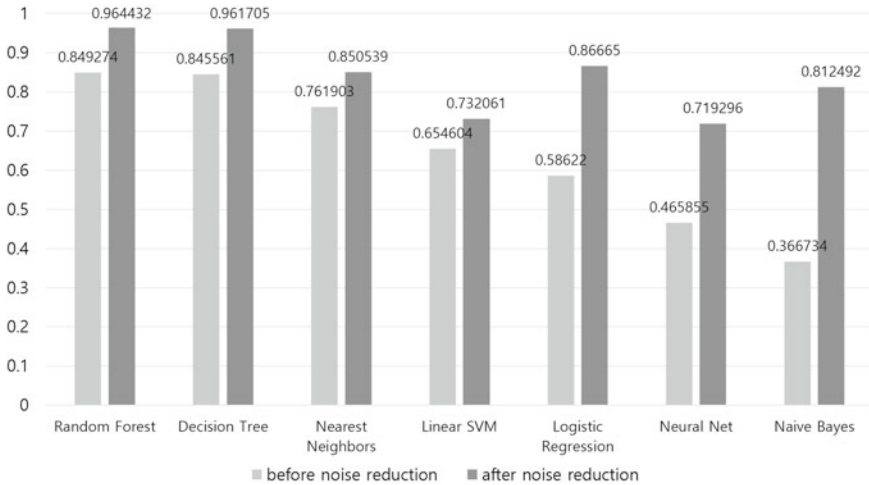


Fig. 1 Attack type classification results (detection accuracy) before and after noise reduction

5 Conclusion

Data with much of cross-references is closely connected with cyber threat incidence that impacts on broad areas, and this means the critical case. Data with low cross-references mean normal network resources that are not related to accidents. In this research, we proposed a new noise reduction algorithm for the CTI dataset. Based on the importance evaluation algorithm about threat information, we distinguished the data with a significant cross-reference rate and data with a low cross-reference rate from the dataset. Our result improved almost 12% of detection performance in cyber threat detection with much small and efficiently filtered dataset. By reducing the complexity of the CTI data through the proposed algorithm, it is expected that the ability to anticipate the attack through the CTI system and preemptively response process to the cyber threat can be significantly improved. In the future study, we have the plan to adapt this importance-evaluation methodology to more elaborate techniques such as STIX and TAXII.

References

1. Gubbi J, Buyya R, Marusic S, Palaniwami M (2013) Internet of things (iot): a vision, architectural elements, and future directions. *Future Generat Comput Syst* 29(7):1645–1660
2. Abu-Nimeh S, Foo E, Fovino IN, Govindarasu M, Morris T (2013) Cyber security of networked critical infrastructures [guest editorial]. *IEEE Netw* 27(1):3–4
3. Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *Siam Rev* 60(2):223–311

4. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Tech rep, Stanford InfoLab
5. Barnum S (2012) Standardizing cyber threat intelligence information with the structured threat information expression (stix). Mitre Corpor 11:1–22
6. Burger EW, Goodman MD, Kampanakis P, Zhu KA (2014) Taxonomy model for cyber threat intelligence information exchange technologies. In: Proceedings of the 2014 ACM workshop on information sharing & collaborative security. ACM, pp 51–60
7. Kim D, Kim HK (2018) Automated dataset generation system for collaborative research of cyber threat intelligence analysis. [arXiv:1811.10050](https://arxiv.org/abs/1811.10050)
8. Sillaber C, Sauerwein C, Mussmann A, Breu R (2016) Data quality challenges and future research directions in threat intelligence sharing practice. In: Proceedings of the 2016 ACM on workshop on information sharing and collaborative security. ACM, pp 65–70
9. Gong S, Cho J, Lee C (2018) A reliability comparison method for osint validity analysis. *IEEE Trans Industr Inf* 14(12):5428–5435
10. Meier R, Scherrer C, Gugelmann D, Lenders V, Vanbever L (2018) Feedrank: a tamper-resistant method for the ranking of cyber threat intelligence feeds. In: 2018 10th international conference on cyber conflict (CyCon). IEEE, pp 321–344
11. Aggarwal CC (2014) Data classification: algorithms and applications. CRC Press

An Improved DBSCAN Method Considering Non-spatial Similarity by Using Min-Hash



Jin Uk Yoon, Byoungwook Kim, and Joon-Min Gil

Abstract In data mining, there are several clustering algorithms that utilize a spatial attribute to group spatial objects on geometric space. However, a spatial object can have a non-spatial attribute as well as spatial attribute, but there are not many clustering algorithms that utilize both the spatial attribute and the non-spatial attribute yet. Jaccard similarity can be used as one of the ways in which similar spatial objects can be grouped by using the non-spatial attribute, but has the problem of higher calculation costs. Therefore, this paper proposes an improved DBSCAN method that utilizes the spatial attribute and non-spatial attribute in DBSCAN to actually cluster more similar Spatial Objects and uses Min-Hash to reduce the cost of calculating Jaccard similarity. The improved DBSCAN method we propose takes into account the similarity of non-spatial attribute in addition by using Min-Hash, when the neighborhood is obtained according to euclidean distance in the rangeQuery of existing DBSCAN process. We use real dataset to compare and analyze the results of classical DBSCAN with that of our method to demonstrate the applicability to real world and we use synthetic dataset composed with various experimental variables to compare performance of using Jaccard similarity with using Min-Hash.

Keywords Clustering · DBSCAN · Non-spatial attribute · Min-Hash

J. U. Yoon

Department of Computer Science, Dongguk University, Gyeongju, South Korea

e-mail: jinuknamja@dongguk.ac.kr

B. Kim

Department of Computer Engineering, Dongguk University, Gyeongju, South Korea

e-mail: bwkim@dongguk.ac.kr

J.-M. Gil (✉)

School of Information Technology Engineering, Daegu Catholic University, Gyeongsan, Korea

e-mail: jmgil@cu.ac.kr

© Springer Nature Singapore Pte Ltd. 2021

J. J. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,

Lecture Notes in Electrical Engineering 715,

https://doi.org/10.1007/978-981-15-9343-7_84

1 Introduction

Clustering is one of the major methods of data mining, the process of finding meaningful knowledge with grouping the objects, which have similar properties in the entire dataset. The object with similar properties are grouped into one subset, called clusters and that in the same cluster are similar to each other, but that in the other clusters are not alike each other. Cluster analysis can be utilized in many areas, including information retrieval [1], text mining [2], spatial database [3, 4], sequence data analysis [5, 6], web data mining [7, 8], DNA analysis [9, 10], network [11, 12], data stream [13]. Clustering methods are mainly categorized into five types [14]: partition-based, hierarchical-based, density-based, grid-based, and model-based. Among those methods, DBSCAN [14], a density-based clustering is a method to clustering spatial objects where is dense in geometric space. The spatial object can be represented points, lines, and polygons in geometric space and used as roads, buildings, rivers, and tourist attractions on a real map.

In DBSCAN, spatial objects are grouped to one cluster within defined radius if it is dense by using spatial attributes (e.g. longitude, latitude) for measuring Euclidean distance. In addition, DBSCAN has the following characteristics: (1) No need to determine the number of clusters in advance. (2) Arbitrary shaped cluster such as linear, concave, oval, etc. (3) Not sensitive to noise compared to other clustering algorithms (e.g. K-means). DBSCAN is being used in spatial clustering, outlier search, hot spot search, pattern analysis, classification, obstacle detect, network.

However, in the real world, the spatial object can not only be represented on a geometric space, but it can also represent a variety of meanings. In five models of the spatial object (geometric, feature, network, alignment, transformation) were introduced. Therefore, it does not mean much to group the spatial objects on geometric space by simply utilizing the spatial attribute alone. To improve this problem, when clustering spatial objects, it is necessary to use spatial attribute and non-spatial attribute together to take into account the similarity and the proximity of spatial objects. Non-spatial attribute is not attribute meaning spatial property in geometric space such as text, temperature, time, etc. Depending on non-spatial attributes, the spatial objects can be represented in a variety of meanings and new ways to leverage these attributes to actually group more similar spatial objects in DBSCAN were also proposed.

However, with the recent increase in demand for spatial objects and the ability to express a variety of attributes, the non-special attribute of spatial objects can also have a huge amount. Because of this, it cause the calculation cost to be very high that the similarity of non-spatial attribute is obtained using Jaccard similarity and this is called curse of dimensionality. Therefore, this paper and propose improved DBSCAN method to cluster more actually similar spatial objects by using spatial attribute and non-spatial attribute together and introduce Min-Hash in improved DBSCAN method to efficiently estimate Jaccard similarity to reduce this calculation cost.

The algorithms proposed in this paper proceed as follows. (1) Generate a limited dimension of subspace for high-dimensional non-spatial attribute through Min-Hash. (2) In clustering phase of DBSCAN, find spatial objects that are similar and adjoin each other by calculating similarity of subspace and measuring Euclidean distance. (3) Create a new cluster if the number of spatial objects meeting the conditions in 2 is higher than the user-defined threshold. (4) If a new cluster is created, the cluster gradually expands by repeating the steps of 2, 3, 4 to spatial objects in the cluster.

2 Related Works

2.1 DBSCAN

DBSCAN [15] is one of the density-based clustering algorithms and requires two parameters: ϵ , Minpts . About some spatial object ρ , ϵ -neighborhood of ρ means all neighborhood within the defined epsilon (ϵ) of ρ and Minpts means at least the number of ϵ -neighborhood that satisfies high density.

The overall process of DBSCAN be summarized into the following steps: (1) Using Euclidean distance, obtain ϵ -neighborhood of a particular spatial object ρ . (2) Create a new cluster if the number of spatial objects in ϵ -neighborhood is higher than Minpts and there are no clusters that already belong. (3) Repeat steps 1 and 2 for spatial objects in ϵ -neighborhood, and gradually expand the cluster. DBSCAN is performed with $O(n^2)$ in a Naïve method, but with spatial indexes such as R-Tree and X-Tree, it is performed a little faster.

DBSCAN has various modifications to solve the existing problems. OPTICS has improved the difficulty of not distinguishing various densities according to the two parameters (ϵ , Minpts) that can form clusters in classical DBSCAN. DENCLUE uses the kernel function of density distribution to find clusters and improve the problem of fixed parameter structure. In large databases, the Incremental DBSCAN improved the problem to update a large number of clustering information when new object is inserted or removed from an existing cluster. SDBDC, distributed DBSCAN, was divided into two stages into local level and global level, conducting DBSCAN separately to reduce unnecessary calculation cost. In recent years, There are a number of studies related to DBSCAN, including to efficiently cluster large amounts of data by using MapReduce or Spark, to reduce run time when clustering with high-dimensional object, to consider the density of clusters be varied.

Also proposed are ways to use non-spatial attribute in DBSCAN. ST-DBSCAN proposed using the location of the spatial object as a spatial attribute and the temperature data as a non-spatial attribute, and measuring the Euclidean distance for the two attributes to cluster the adjacent spatial objects. DBSTexC proposed a method of clustering the spatial objects that have related text based the POI position by using text data that have information about POI (point-of-Interest) and location of user on Twitter. Clustering Public Transit Stops using an Improved DBSCAN Algorithm further considered the similarity of bus stop names through editing distance, and proposed a method for clustering public transportation stops related to each other on the map. GDBSCAN generalizes the definitions of ϵ -neighborhood and Minpts, introducing different definition and clustering methods according to the various non-spatial attributes defined at a certain situations.

2.2 Min-Hash

Min-Hash is a type of Local Sensitive Hashing (LSH) technique that use random permutation to estimate the similarity of two sets of Jaccard similarity. To perform random permutation of two sets of elements, the uniform distribution hash function h is need to used and the minimum value h_{\min} of result values that elements of two set is mapped from hash function, called Min-Hash Value is used to estimate Jaccard similarity of two sets. The probability that the Min-Hash value of the two sets A and B being compared are same each other is Jaccard similarity of two sets A and B.

$$\Pr[h_{\min}(A) = h_{\min}(B)] = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

The process for estimating the Jaccard similarity between two sets A and B through Min-Hash is as follows: (1) Select the k hash functions that maps the members of union A and B to distinct integers. (2) Find k Min-Hash Values, the minimum value h_{\min} of result values that elements mapped from each hash function. (3) Obtain k Min-Hash Values and make these as a Subspace of size k (this is called a signature). (4) Compare the i th ($1 \leq i \leq k$) Min-Hash Value between the two signatures and count the number of the same Min-Hash Value each other. (5) Estimate Jaccard similarity of two sets A and B by counting the number of equal Min-Hash Values each other on a subspace with total size k (Fig. 1).

Min-Hash is widely applied, including anti-plagiarism, graph or image analysis and meta-genetic analysis, in particular, as a way to efficiently estimate similarities in numerous data areas that can be expressed in set. Min-Hash was applied at first to calculate the degree of similarity between the two large documents under comparison. To calculate the similarity, the large document is represented with q -grams of huge amount and q -grams about the document is converted as small data through Min-Hash so that the similarity between the two documents can be measured efficiently. In proposed is method that the similarity between graphs could be effectively

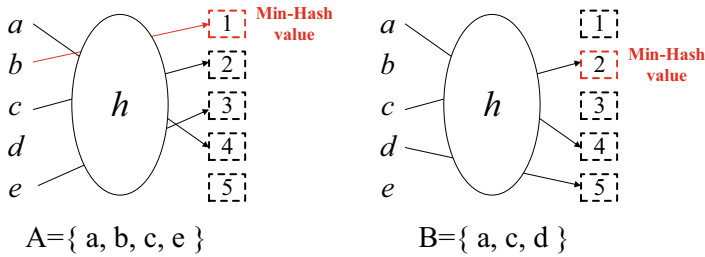


Fig. 1 Min-hash value

calculated by using Min-Hash in the process of comparing them by dividing them into subgraphs. In proposed is method, in large image databases to store not image object with high-dimensional vectors but low-dimensional vector and increase search efficiency by using Min-Hash. In showed that Min-Hash performs very well when the set size are similar each other and significantly decreases in performance when they are different in size, proposing a new effective improved Min-Hash method to detect the appearance and absence of microorganisms in the metagenomic dataset.

In this paper, we introduced Min-Hash in DBSCAN and implemented improved DBSCAN using spatial attribute and non-spatial attribute of spatial object. By using Min-Hash, we can calculate the Jaccard similarity between high-dimensional non-spatial attributes. In next section, we describe how to estimate the similarity of non-spatial attribute through the Min-Hash and what the process of improved DBSCAN.

3 Algorithm

In this section, we introduce two algorithms of the improved DBSCAN method considering the similarity of non-spatial attributes in the classical DBSCAN. The first algorithm use Jaccard similarity and the second algorithm use Min-Hash we proposed. Both algorithms have $Minpts$, ϵ , τ as parameters, but the improved DBSCAN using Min-Hash additionally uses the size of signature k .

Algorithm 1. An Improved DBSCAN (Jaccard similarity)

Input: A set \mathcal{D} of \mathcal{T} tuples, $Minpts$, ε , τ

Output: A set \mathcal{D}' of \mathcal{T} labeled tuples

1. cluster id $C \leftarrow 0$
2. **for each** unclassified tuple $\mathcal{T} \in \mathcal{D}$ **do**
3. $N_{(\varepsilon, \tau)}(\mathcal{T}) \leftarrow RangeQuery(\mathcal{T}, \varepsilon, \tau)$
4. **if** $|N_{(\varepsilon, \tau)}(\mathcal{T})| \geq Minpts$ **then**
5. Set \mathcal{T} cluster id to C
6. $ExpandCluster(\mathcal{T}, N_{(\varepsilon, \tau)}(\mathcal{T}), C, \varepsilon, \tau, Minpts)$
7. $C \leftarrow C + 1$
8. **else**
9. Label \mathcal{T} as Noise

A random tuple \mathcal{T} is selected from the entire data set \mathcal{D} , and the neighborhood of the tuple \mathcal{T} is stored in $N_{(\varepsilon, \tau)}(\mathcal{T})$ through RangeQuery. If the number of tuple stored in $N_{(\varepsilon, \tau)}(\mathcal{T})$ is more than $Minpts$, tuple \mathcal{T} becomes core and creates a new cluster centering around neighborhood of \mathcal{T} and calls ExpandCluster to expand the cluster.

4 Conclusions

In this paper, we propose an improved DBSCAN algorithm that can consider non-spatial attributes additionally as a factor for clustering spatial objects in geographic space and Min-Hash could be used for spatial objects with high-dimensional non-spatial attribute. We will mainly use image and text data among real data so as to analyze various result of clustering. Also, we will study various clustering algorithms that can consider the diversity of spatial object as well as DBSCAN in the future.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1F1A1062039).

References

1. Hotho A, Nürnberger A, Paaß G (2005) A brief survey of text mining
2. Chandra E, Anuradha VP (2011) A survey on clustering algorithms for data in spatial database management systems. Int J Comput Appl
3. Kolatch E (2001) Clustering algorithms for spatial databases: a survey
4. Guralnik V, Karypis G (2001) A scalable algorithm for clustering sequential data. In: Proceedings 2001 IEEE international conference on data mining

5. Ferreira D, Zacarias M, Malheiros M, Ferreira P (2007) Approaching process mining with sequence clustering: experiments and findings. In: International conference on business process management
6. Nirkhi S, Hande K (2004) A survey on clustering algorithms for web applications. In: Proceedings of the 2008 international conference on semantic web & web services
7. Vakali A, Pokorny J, Dalamagas T (2004) An overview of web data clustering practices. In: International conference on extending database technology
8. Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, Brown P, Haas BJ, Salzberg SL (1999) Clustering methods for the analysis of DNA microarray data. Educational Technology Publications
9. Volfovsky N, Haas BJ, Salzberg SL (2000) A clustering method for repeat analysis in DNA sequences. Educational Technology Publications
10. Abbasi AA, Younis M (2007) A survey on clustering algorithms for wireless sensor networks. *Comput Commun*
11. Afsara MM, Mohammad H, Tayarani N (2014) Clustering in sensor networks: a literature survey. *J Netw Comput Appl*
12. Sharma N, Masih S, Makhija P (2018) A survey on clustering algorithms for data streams. *Int J Comput Appl*
13. Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya A, Fofou S, Bouras A (2014) A survey of clustering algorithms for big data: taxonomy and empirical analysis. *IEEE Trans Emerg Topics Comput* 2(3):267–279
14. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD-96 proceedings
15. Lim S, Lee S, Kim S-C, Clustering of detected targets using DBSCAN in automotive radar systems. In: 2018 19th international radar symposium (IRS)