Studies in Computational Intelligence 850

Roger Lee *Editor*

Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing

Springer

Studies in Computational Intelligence

Volume 850

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are submitted to indexing to Web of Science, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink.

More information about this series at http://www.springer.com/series/7092

Roger Lee
Editor

Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing

Springer

Editor
Roger Lee
Software Engineering and Information
Technology Institute
Central Michigan University
Mt. Pleasant, MI, USA

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-030-26427-7 ISBN 978-3-030-26428-4 (eBook)
https://doi.org/10.1007/978-3-030-26428-4

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

Foreword

The purpose of the 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2019) held on July 8–10, 2019 in Toyama, Japan is aimed at bringing together researchers and scientists, businessmen and entrepreneurs, teachers and students to discuss the numerous fields of computer science, and to share ideas and information in a meaningful way. This publication captures 16 of the conference's most promising papers, and we impatiently await the important contributions that we know these authors will bring to the field.

In chapter "Generating Linear Temporal Logics Based on Property Specification Templates", Weibin Luo, Hironori Washizaki and Yoshiaki Fukazaw introduce a method to create linear temporal logics (LTLs) based on property specification patterns. They experimentally compare the required time and accuracy of their approach to those using property specification patterns. Their approach can improve the creation of LTLs in terms of speed and accuracy.

In chapter "Rule-Based BCG Matrix for Product Portfolio Analysis", Chih-Chung Chiu and Kuo-Sui Lin propose a new Boston Consulting Group (BCG) method, which is a dynamic trend analysis method, and different time frames are considered for product portfolio analysis. A numerical case study was conducted to demonstrate that the proposed BCG method has achieved the objectives of the study for company products' portfolio analysis.

In chapter "New Cost-Consequence FMEA Model for Information Risk Management of Safe And Secure SCADA Systems", Kuo-Sui Lin proposes a new cost-consequence Failure Mode and Effects Analysis (FMEA) model. It not only can recover traditional RPN-based FMEA problems but also can evaluate, prioritize, and correct safety and security of a SCADA system's failure modes.

In chapter "A Dynamic Privacy Preserving Authentication Protocol in VANET Using Social Network", Syed Asad Shah, Chen Gongliang, Li Jianhua and Yasir Glani propose a dynamic privacy-preserving authentication protocol using a hybrid combination of symmetric advanced encryption standard (AES) and asymmetric elliptic curve cryptography (ECC). They also propose a dynamic region topology

algorithm to ensure the message gets delivered to desired vehicles in a region using the social network.

In chapter "A Negotiation Strategy Based on Compromising Degree", Shun Okuhara and Takayuki Ito propose an explainable concession process using a constraint relaxation process. The authors also describe methods min, distance, and distance min, random for the size of the removed constraint. Experimental results demonstrate that distance strategies are effective.

In chapter "Software Developer Performance Measurement Based on Code Smells in Distributed Version Control System", Natach Jongprasit and Twittie Senivongse propose a method and a supporting tool for measuring the performance of individual software developers in a Git project based on code smells.

In chapter "Decision Support System for Choosing an Elective Course Using Naive Bayes Classifier", Abiyoga, Arya Wicaksana and Ni Made Satvika Iswari propose a decision support system to assist the students in choosing an elective course based on not only their interest but also academic skills. The system uses Naive Bayes classifier and Laplace smoothing for the classification process.

In chapter "Multilevel Video Access Control Mechanism Using Low Power Based Audio Watermarking", Sunyoung Choi, Youngmo Kim and Ung-Mo Kim propose a multilevel video access control mechanism to prevent leakage of personal information in security video.

In chapter "A Study of Persona Research on Domestic Music Applications", HaeKyung Chung and JangHyok Ko conduct a literature study to understand overall understanding of mobile music service and complete persona through survey methods such as surveys and in-depth interviews. The goal was to design an interface aspect that requires a simple yet intuitive design.

In chapter "Analysis of Large-Scale Diabetic Retinopathy Datasets Using Texture and Blood Vessel Features", Devvi Sarwinda, Ari Wibisono, Hanifa Arrumaisha, Zaki Raihan, Rosa N. Rizky FT, Rico Putra Pradana, Mohammad Aulia Hafidh and Petrus Mursanto classify the stages of diabetic retinopathy using a large-scale dataset that consists of 35,126 fundus images. The classification of diabetic retinopathy includes five stages, from normal to proliferative diabetic retinopathy.

In chapter "Combination of 1D CNN and 2D CNN to Evaluate the Attractiveness of Display Image Advertisement and CTR Prediction", Wee Lorn Jhinn, Poo Kuan Hoong and Hiang-Kwang Chua propose a method to evaluate and analyze the elements of attractiveness within the display advertisement in Facebook Advertisement platform by applying the 2D CNN on the display advertisement images while 1D CNN on the click metric data, respectively.

In chapter "Functional Reactive EDSL with Asynchronous Execution for Resource-Constrained Embedded Systems", Sheng Wang and Takuo Watanabe present a functional reactive embedded domain-specific language (EDSL) for resource-constrained embedded systems and its efficient execution method.

In chapter "Adaptive Midpoint Relay Selection: Enhancing Throughput in D2D Communications", Ushik Shrestha Khwakhali, Prapun Suksompong and Steven Gordon propose adaptive midpoint relay selection scheme that is designed to

achieve the performance of MRSS-ST when the social trust among the nodes are low and achieve the performance of M-nearest when the social trust are high.

In chapter "Collaborative SCM System for Sustainability in the Manufacturing Supply Chain", Donghyuk Jo defines the success of export-based small and medium manufacturing firms as export performance, and verifies the effects of the collaborative SCM activities of firms and the establishment of the supply chain on the supply chain performance and export performance. The purpose of this study is to identify the importance of collaborative SCM activities to enhance the competitiveness of export-based small and medium manufacturing firms and to suggest strategic directions of SCM.

In chapter "A Study on the Effect of Cultural Capital on the Innovative Behavior", Hye Jung Kim, Jongwoo Park and Myeong Sook Park provide new perspective by the principle of individual innovation from the perspective of humanitarianism based on the theoretical background that individuals constituting an organization, not the object of the organizational unit, should be the principal agent of corporate innovation to succeed in corporate innovation. Purpose of this study is to provide theoretical framework and implications needed for the strategic innovation measures, improvement of management techniques, and development of new operational management model in the rapidly changing era.

In chapter "Evaluation of Technology Transfer Performance in Technology-Based Firms", Donghyuk Jo and Jongwoo Park aim to determine the factors affecting the performance of the firms with transferred public technology and suggest implications thereof. To that end, this study empirically analyzed the effects of transferred technology value (technology transaction amount, technology readiness level) and absorptive capacity (potential absorption capability, feasible absorption capacity) on technology transfer performance.

It is our sincere hope that this volume provides stimulation and inspiration, and that it will be used as a foundation for works to come.

July 2019

Takayuki Ito
Nagoya Institute of Technology
Nagoya, Japan

Contents

Contributors

Abiyoga Department of Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

Hanifa Arrumaisha Undergraduate Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Chih-Chung Chiu Department of Information Management, Aletheia University, New Taipei City, Taiwan (R.O.C.)

Sunyoung Choi College of Software, Sungkyunkwan University, Suwon, Republic of Korea

Hiang-Kwang Chua Axiata Digital Advertising (ADA), Kuala Lumpur, Wilayah Persekutuan, Malaysia

HaeKyung Chung Department of Visual Communication and Media Design, Konkuk University, Chungju-si, Republic of Korea

Yoshiaki Fukazawa School of Science and Engineering, Waseda University, Tokyo, Japan

Yasir Glani School of Information Security Engineering, Shanghai Jiao Tong university, Shanghai, China

Chen Gongliang School of Information Security Engineering, Shanghai Jiao Tong university, Shanghai, China

Steven Gordon School of Engineering and Technology, CQUniversity, Rockhampton, Australia

Mohammad Aulia Hafidh Undergraduate Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Poo Kuan Hoong Axiata Digital Advertising (ADA), Kuala Lumpur, Wilayah Persekutuan, Malaysia

Ni Made Satvika Iswari Department of Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

Takayuki Ito Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Aichi, Japan

Wee Lorn Jhinn Axiata Digital Advertising (ADA), Kuala Lumpur, Wilayah Persekutuan, Malaysia

Li Jianhua School of Information Security Engineering, Shanghai Jiao Tong university, Shanghai, China

Donghyuk Jo Department of Business Administration, Soongsil University, Seoul, South Korea

Natach Jongprasit Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

Hye Jung Kim Department of Business Administration, Soongsil University, Seoul, South Korea

Ung-Mo Kim College of Software, Sungkyunkwan University, Suwon, Republic of Korea

Youngmo Kim Department of Computer Science and Engineering, Soongsil University, Seoul, Republic of Korea

JangHyok Ko Division of Computer and Mechatronics, Sahmyook University, Seoul, Republic of Korea

Kuo-Sui Lin Department of Information Management, Aletheia University, New Taipei City, Taiwan (R.O.C.)

Weibin Luo School of Science and Engineering, Waseda University, Tokyo, Japan

Petrus Mursanto Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Shun Okuhara School of Medical Sciences, Fujita Health University, Toyoake, Aichi, Japan

Jongwoo Park Department of Business Administration, Soongsil University, Seoul, South Korea

Myeong Sook Park Department of Business Administration, Soongsil University, Seoul, South Korea

Rico Putra Pradana Undergraduate Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Zaki Raihan Undergraduate Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Rosa N. Rizky FT Undergraduate Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Devvi Sarwinda Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia

Twittie Senivongse Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

Syed Asad Shah School of Information Security Engineering, Shanghai Jiao Tong university, Shanghai, China

Ushik Shrestha Khwakhali Vincent Mary School of Engineering, Assumption University, Samut Prakan, Thailand;
School of ICT, SIIT, Thammasat University, Bangkok, Thailand

Prapun Suksompong School of ICT, SIIT, Thammasat University, Bangkok, Thailand

Sheng Wang Department of Computer Science, Tokyo Institute of Technology, Meguroku, Tokyo, Japan

Hironori Washizaki School of Science and Engineering, Waseda University, Tokyo, Japan;
National Institute of Informatics, Tokyo, Japan;
SYSTEM INFORMATION CO., LTD, Tokyo, Japan;
eXmotion Co., Ltd, Tokyo, Japan

Takuo Watanabe Department of Computer Science, Tokyo Institute of Technology, Meguroku, Tokyo, Japan

Ari Wibisono Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Arya Wicaksana Department of Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

Generating Linear Temporal Logics Based on Property Specification Templates

Weibin Luo, Hironori Washizaki and Yoshiaki Fukazawa

Abstract Temporal logics are widely used in software verification such as model checking. However, creating temporal logics such as linear temporal logics (LTLs) based on property specifications written in a natural language is difficult due to practitioners' unfamiliarity with property specifications and notations of temporal logics. Although property specification patterns have been introduced to help write correct temporal logics, creating temporal logics using property specification patterns requires an understanding of the pattern system. Since some patterns are difficult to understand, especially for beginners, and the final temporal logics are usually complicated, creating temporal logics using pattern systems is time consuming and error-prone. Here, we introduce a method to create LTLs based on property specification patterns. We experimentally compare the required time and accuracy of our approach to those using property specification patterns. Our approach can improve the creation of LTLs in terms of speed and accuracy. Although our experiment is implemented in Japanese, the results should be applicable to other languages such as English. We also provide a visualization scheme so that practitioners can understand the generated LTLs and confirm that they are correct.

W. Luo · H. Washizaki (✉) · Y. Fukazawa
School of Science and Engineering, Waseda University, Tokyo, Japan
e-mail: washizaki@waseda.jp

W. Luo
e-mail: luoweibinb@moegi.waseda.jp

Y. Fukazawa
e-mail: fukazawa@waseda.jp

H. Washizaki
National Institute of Informatics, Tokyo, Japan

SYSTEM INFORMATION CO., LTD, Tokyo, Japan

eXmotion Co., Ltd, Tokyo, Japan

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_1

1 Introduction

Software reliability has become an important issue in the software development process. Often reliability is verified using model checking. Model checking [1] is a formal method to determine if a model satisfies the given property specifications, which are usually written in a natural language. In model checking, the system properties are translated as temporal logic formulas. Then the model checker verifies if the system is consistent with the given properties. Temporal logics such as linear temporal logic (LTL) [2], computation tree logic (CTL) [3], and action computation tree logic (ACTL) [4] are commonly used in model checking. However, users must have a solid mathematical understanding to create these temporal logics from property specifications because temporal logics are difficult to understand [5]. To address this problem, many property patterns [6–9] have been proposed to make temporal logics easier to understand and create.

Dwyer et al. [6, 7] proposed a pattern system, which consists of patterns, scopes, and pattern mapping templates. And, there are some approaches for automatic verification utilizing the pattern system and templates [10, 11]. Patterns describe what must occur. Scopes detail when the patterns must hold. Pattern mapping templates plot the patterns and scopes to temporal logics with state formulas. Practitioners select the appropriate pattern and scope to match the given properties. Then they use pattern mapping to obtain a template of the temporal logic such as LTL. Afterwards, practitioners can create real temporal logics by replacing the state formulas with real states. However, this method has two problems. First, users must understand the whole pattern system (all patterns and scopes). Some patterns are too similar to distinguish (e.g., Global Response pattern and After Existence pattern). Hence, creating the temporal logic is time consuming. Second, since users have to replace the states manually, the replacement process is error-prone when states and templates are complicated.

In this paper, we propose a method to automatically generate real temporal logics from the given property specifications based on the above property patterns. Instead of selecting the pattern and scope, users refine the original property specification to match the prepared specification template. Then our system generates the real temporal logic based on the refined specification. We also propose a pattern-based approach, Property Specification Graph (PSG), to visualize the generated temporal logics. PSG help practitioners verify the correctness of the generated temporal logic.

To evaluate our approach, we conducted an experiment that compares the proposed approach with the existing method in which temporal logics are created using the property pattern system. In the experiment, participants were asked to create LTLs using both methods. The experiment addressed the following research questions:

RQ1. Does our approach improve the speed of creating temporal logics?
RQ2. Does our approach improve the accuracy of creating temporal logics?
RQ3. What are the advantages and the disadvantages of the proposed approach compared to the traditional method?

The proposed approach can help create LTLs in terms of speed and accuracy. Furthermore, 70% of the participants felt that the proposed method is more efficient than the traditional method. 90% reported that the traditional method is difficult to use when selecting the appropriate pattern and replacing the states in the template, whereas 40% said that the specification templates in the proposed method are difficult to understand and use.

The remainder of the paper is organized as follows. Section 2 overviews the property pattern system. Section 3 introduces our approach to generate temporal logics based on specification templates. Section 4 describes the experiment, and Sect. 5 compares our approach with the traditional method. Section 6 applies our approach to a real-world specification as an example. Section 7 discusses the threats to validity and Sect. 8 describes related work. Finally, Sect. 9 concludes the paper and provides future work.

2 Background

Dwyer et al. [7] collected over 500 examples of property specifications for finite-state verification tools, and found that nearly 92% of the properties can be classified into a hierarchy of basic patterns based on semantics. Each pattern is comprised of an intent, scope, and mapping template. The intent describes the structure of the behavior. The scope constrains the extent of program execution over which the pattern must hold. The mapping template plots common specification formalisms (LTL, CTL and QRE), examples of known uses, and relationships to other patterns. For example, the Absence pattern is for the portion of a system's execution that is free of certain events or states. Below we briefly describe the patterns, scopes, and template mappings in the pattern system.

Patterns: Dwyer et al. [7] introduced eight basic patterns:
Absence: A given state/event does not occur within the scope.
Existence: A given state/event must occur within the scope.
Bounded Existence: A given state/event must occur k times within the scope.
Universality: A given state/event occurs throughout the scope.
Precedence: State/event P must always be preceded by state/event Q within a scope.
Response: State/event P must always be followed by state/event Q within a scope.
Chain Precedence: A sequence of states/events $P_1, \ldots, P_n$ must always be preceded by a sequence of states/events $Q_1, \ldots, Q_m$.
Chain Response: A sequence of states/events $P_1, \ldots, P_n$ must always be followed by a sequence of states/events $Q_1, \ldots, Q_m$.

Scopes: There are five types:
Global: The pattern must hold during the entire program execution.
Before: The pattern must hold up to a given state/event.
After: The pattern must hold after the occurrence of a given event/state.

Between: The pattern must hold from a given state/event Q to another given state/event R.
After-Until: Similar to Between, but the designated part of the execution continues even if the second state/event does not occur.

Mapping Templates: Each scope has a mapping template, which plots the system properties to specification formalisms (LTL, CTL, or QRE) based on practitioners' selection of patterns and scopes. Below are two examples of template mappings of LTLs:

Absence with a ***Global*** scope:
Explanation: Globally, P is false.
Template mapping: $\Box(!P)$

Response with ***Between*** scope:
Explanation: Between Q and R, S responds to P.
Template mapping: $\Box((Q \,\&\, !R \,\&\, \Diamond R) \rightarrow (P \rightarrow (!R \,U\, (S \,\&\, !R))) \,U\, R)$

The following steps are used to create a temporal logic. (1) Practitioners select the pattern and scope based on their understanding of the original property specifications. (2) Practitioners replace state formulas with real states to obtain the real temporal logic.

This approach to create temporal logics has several problems. (1) Practitioners must choose the appropriate patterns and scopes, which requires a full understanding of the pattern system. Some definitions such as between and after-until are similar, making them difficult to distinguish, especially for beginners. (2) Some template mappings are simple (e.g., Absence/Global), while others (e.g., Response/Between) are extremely complicated. Replacing state formulas with real states in these complicated template mappings can be error-prone. (3) Since the original property specifications are typically written in a natural language and the real states in the final temporal logics are real parameters (usually variables), practitioners may have trouble converting a natural language to real states.

3 Approach

We introduce a new approach to create temporal logics based on the property pattern system. Our approach focuses on creating LTLs since they are relatively simple and easy to understand compared to other temporal logics such as CTL or QRE. Figure 1 overviews our approach. It should be noted that the original study was implemented in Japanese.

Our approach is divided into three steps. (1) Extract the mapping table. (2) Refine the property specifications. (3) Generate LTL and PSG. Below each step is explained in detail.

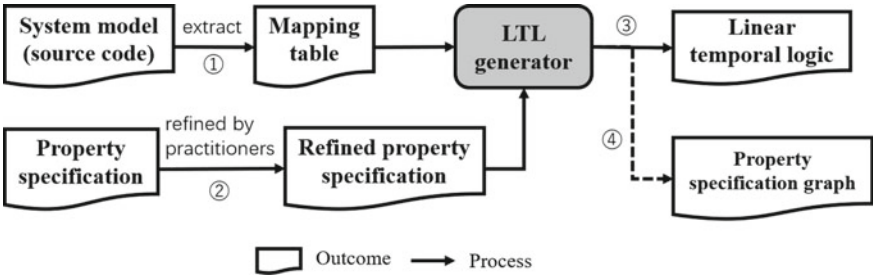


Fig. 1 Overview of our approach

3.1 Extract the Mapping Table

Since the original property specifications are written in a natural language and the final LTL contains the relationships among real parameters (variables), the connections between the parameters in the original property specifications and the variables in the system model (source code) must be established. In the first step, practitioners create a mapping table by connecting the parameters in a natural language with the variables in the source code. Each system model should have only one mapping table. This table is then used to automatically convert the refined property specifications into a real LTL in step 3.

3.2 Refine the Property Specifications

Specification writers do not always follow a certain pattern system when writing property specifications. To bridge this gap, we introduce a set of property specification templates based on the property pattern system. Each pattern/scope has one property specification template as shown below. Currently, our specification templates do not contain Chain Precedence/Response patterns because these are too complicated and are rarely used (9 of 555 cases) [7]. However, we plan to add these two patterns in the future.

Patterns:
Absence: (P) must not hold.
Existence: (P) must hold at least once.
Bounded Existence: (P) must hold at least [k] times.
Universality: (P) must hold all the time.
Precedence: (P) must hold if the subsequent (S) holds.
Response: If (P) holds, then the subsequent (S) must hold.

Scopes:
Global: Globally

Before: Before (Q) holds
After: After (Q) holds
Between: After (Q) holds and before (R) holds
After-Until: After (Q) holds and until (R) holds

In step 2, practitioners refine the original property specifications using property specification templates. Specifically, they choose one template from the patterns and one from the scopes. Then they replace the state formulas (P, S, R, Q) with parameters in the original property specifications. Notice that practitioners do not have to have knowledge about the pattern system (e.g., what does precedence mean?). They only see the property specification templates. Additionally, the refined property specifications only consist of a natural language without any source code level parameters.

3.3 Generate LTL

Based on the mapping table and refined property specifications, the LTL generator generates a real LTL. That is, the LTL generator selects the corresponding template mapping based on the property specification template selected by practitioners. Then it replaces the parameters with variables in the source code according to the mapping table.

3.4 Generate PSG

The LTL generator also creates PSG, which is used to help practitioners better understand LTL and check if the generated LTL is correct. Similar to the property specification templates, each pattern/scope has a PSG template, and the combination of a PSG pattern template and a PSG scope template is the complete PSG.

The patterns in the PSG template consist of three states and two state transition arrows (Fig. 2). Below each pattern is discussed.

Absence: This pattern describes the case where state P must be false throughout the scope, the three states should be state P with an exclamation mark before it (Fig. 2a).
Universality: This pattern is the opposite of Absence. We simply remove the exclamation mark before each state P (Fig. 2b).
Existence: This pattern indicates that state P must hold at least once. Thus, one of the three states should be state P (Fig. 2c). However, a circle without a state formula means that it is not specific and can be any state.
Bounded existence: Similar to *Existence*, *Bounded existence* adds a quantifier *k* to imply that state P must hold at least k times (Fig. 2d).
Response: We use two states, *P* and *S*, to describe that state S responds to state P. In Fig. 2e, the asterisk after *P* stands for "if state *P* holds".

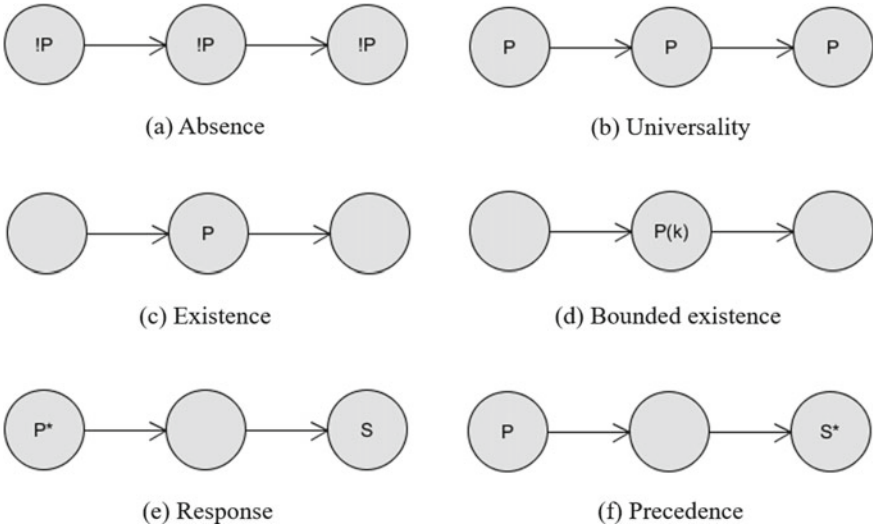


Fig. 2 Property specification graph (patterns)

Precedence: This pattern is similar to *Response*. Figure 2f describes "if state S holds, then state P must hold before."

The scopes of the PSG template consist of scope indicators, states, and state transition arrows. When combining the scope with the patterns, the states with a shadow should be replaced by the pattern in the PSG template. Each scope in the PSG template describes below:

Global: The global scope (Fig. 3a) does not require additional scope indicators. It is the same as the pattern in the PSG template.

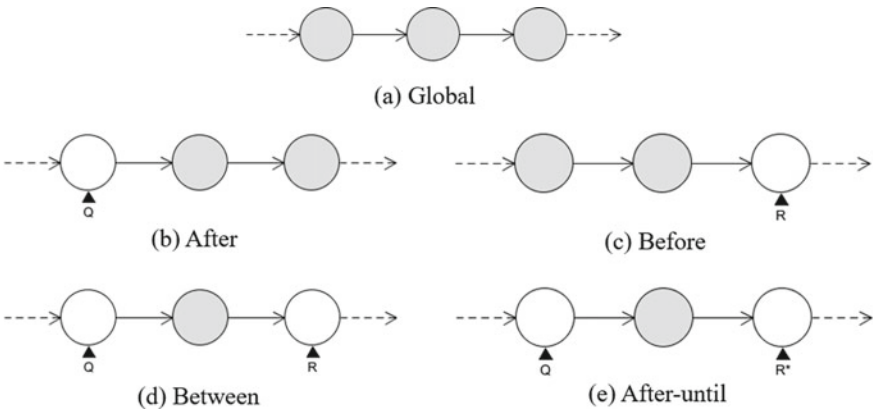


Fig. 3 Property specification graph (scopes)

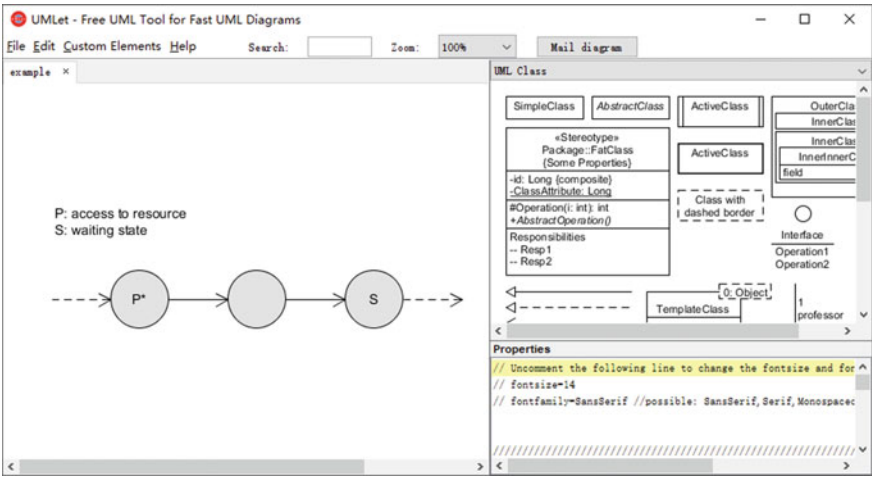


Fig. 4 PSG with UMLet

After: The after scope occurs once a specific situation is occurs. A scope indicator is added to the first state (Fig. 3b).
Before: Similar to *After*, a scope indicator is added. However, it is added to the last state to describe the before scope (Fig. 3c).
Between: The between scope is a combination of *after* and *before*. Thus, two indicators are added to the first and the last states (Fig. 3d).
After-until: This scope differs slightly from *Between*. An asterisk is added to the second scope indicator (Fig. 3e), which means "if state R holds". The second indicator can be omitted if R does not hold.

All PSGs are implemented based on a free UML tool for fast UML diagrams called UMLet [12]. Our PSG only contains basic components (i.e., circles, arrows, and triangles), making it easier to implement on different tools. Figure 4 shows a PSG with UMLet.

4 Example

In this section, we present an example where our approach is used on a realistic OSEK/VDX operating system [13]. OSEK/VDX is a development standard for automobiles that has been widely adopted by manufacturers to develop automotive systems. We extracted the property specifications from the OSEK/VDX operating system specification [13] as follows:

Access to resources never results in a waiting state.

Table 1 Example of a mapping table

Parameter	Variable
Access to resources	access_resource
Waiting state	state_waiting

To apply our approach to this property, practitioners should initially create a mapping table for the whole system. In this step, the mapping table relates to the property specifications shown in Table 1. Parameter *access to resources* is related to the variable *access_resource*, while parameter *waiting state* refers to the variable *state_waiting*. Notice that the variable name depends on the implemented system in the real world. It should also be noted that only two relationships are listed here, but the whole mapping table should be a large set of all connections.

Next, practitioner refines the original property specifications by choosing the most appropriate property specification template. In this case, a correct answer might be:

Globally, if (*access to resources*) holds, then the subsequent (!*waiting state*) must hold.

In this step, several points should be considered. 1. After selecting the appropriate property specification templates, practitioners should replace the state formulas strictly according to the mapping table. Otherwise the LTL generator will not replace the parameters with the corresponding variables. 2. The exclamation mark before the second parameter indicates that the property should be a negative proposition of *waiting state*.

Based on the mapping table and the refined property specifications, the LTL generator generates the following LTL as well as PSG shown in Fig. 4.

$\square$(access_resource $\rightarrow$ $\Diamond$(!state_waiting))

5 Experiment

To evaluate our approach, we conducted an experiment to compare the proposed approach to the existing method (creating temporal logics using the property pattern system). We asked ten participants to create eight LTLs using both our approach and the existing method (four LTLs for each method). The eight questions (Q1 ~ Q8) were divided into two sets where the corresponding questions in each set had similar levels of difficulty. For example, Q1 and Q5 were both about absence patterns with global scopes. These were the simplest questions. Q3 and Q7 were response patterns with global scopes, and were the most complicated questions. Participants were divided into two groups. Group 1 used the existing method first and then the proposed method to create LTLs. Group 2 created LTLs in the reverse order as Group 1.

Since creating a *mapping table* requires an understanding of the source code and the *mapping table* should be created before generating LTLs, we prepared the mapping table before the experiment. Hence, the participants focused on refining the property specifications. We also asked the participants to record the required time to

create each LTL. After the experiment, we asked each participant to answer a questionnaire, which contained the questions shown in Table 4. Through the experiment, we addressed three research questions mentioned in Sect. 1.

All contents of the experiment were in Japanese, including the questionnaire. We translated the results into English. Moreover, the experiment did not include PSG because PSG is implemented after the experiment.

6 Discussion

6.1 *Does Our Approach Improve the Speed of Creating Temporal Logics?*

In the experiment, participants were asked to record the required time (in seconds) to create each LTL. Figure 5 graphs the required times. The required time of the proposed method is shorter than that of the existing method for group 1. The difference is larger when the LTL is complicated (Q3 and Q7, Q4 and Q8). The average required time using the proposed method is 39% faster than that of the existing method. Hence, the proposed method has a smaller required time and time cost.

The average time cost and the p-value of t-test for both are shown in Table 2. The average time cost for our method is lower than the existing one. The p-value of group 1 is less than 0.05, which means that there is a significant difference between the two methods. Because the p-value of group 2 is much larger than 0.05, we cannot

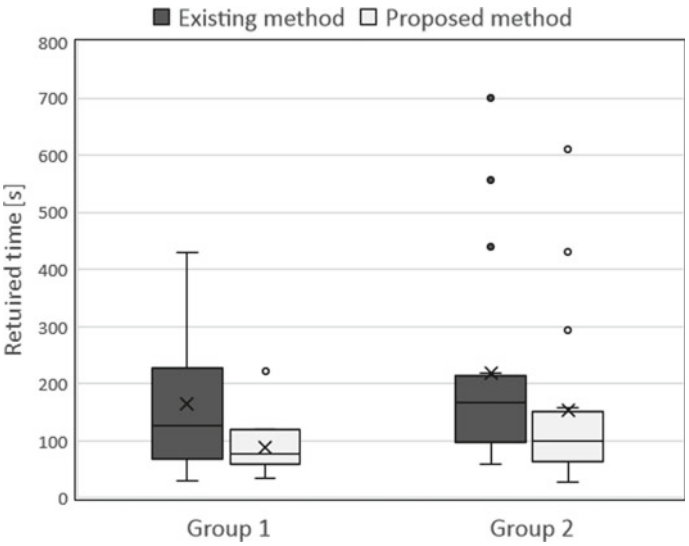


Fig. 5 Comparison of required time

Table 2 Average required time and *p*-value (in seconds) of both groups

	Existing method	Proposed method	*p*-value
G1	164.25	87.63	0.004
G2	219.69	154.13	0.298

conclusively state that there is a significant difference between the two methods. We believe that the participants who used the proposed method first might be familiar with creating LTLs. Therefore, for the existing method, the time cost of group 2 is lower than group 1.

In summary, our method can help to lower the time cost of creating LTLs compared to the existing method. The effect becomes larger when the LTLs are complicated.

6.2 Does Our Approach Improve the Accuracy of Creating Temporal Logics?

Table 3 shows the correctness of each question. The results are mixed. Participants C, D, E, and F failed to write the correct answer in Q3 but created the correct LTL in Q7. Participants B and I found the right answer using the existing method, but failed to do so when using the proposed method. Although there is not a visible trend of accuracy, our approach had an overall total correctness that is 10% higher than the existing method.

6.3 What Are the Advantages and Disadvantages of the Proposed Approach Compared to the Traditional Method?

Table 4 summarizes the questionnaire results by grouping similar answers together. 90% of the participants indicated that it is difficult to select appropriate patterns and replace the states in the template. 40% reported difficulty selecting a specification template. We think that the specification template is hard to understand and should be improved. In question 3, 70% of the participants felt that the proposed method is faster, while 40% reported that the proposed method does not require an understanding of the pattern system. In the last question, participants provided not only ideas to improve our template, but also suggested additional features such as a verification to confirm the generated LTL is correct.

Table 3 Results of each question (◯: correct, ×: incorrect)

		Existing method				Proposed method			
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
G1	A	◯	◯	×	◯	◯	◯	×	◯
	B	◯	◯	◯	◯	◯	◯	×	◯
	C	◯	◯	×	◯	◯	◯	◯	◯
	D	◯	×	×	◯	◯	◯	◯	◯
	E	◯	×	×	×	◯	◯	◯	×
	F	◯	◯	×	◯	◯	◯	◯	◯
G2	G	◯	◯	×	◯	◯	◯	×	◯
	H	◯	◯	◯	◯	◯	×	◯	◯
	I	◯	◯	◯	◯	◯	◯	×	×
	J	◯	×	×	×	◯	◯	×	◯
Total		10	7	3	8	10	9	5	8
		28 (70%)				32 (80%)			

Table 4 Results of the questionnaire

1. **What do you think is the hardest part when creating LTL using the existing method?**
Selecting the pattern and scope (4 of 10)
Replacing the states in the mapping template with the real variables (5 of 10)
2. **What do you think is the hardest part when creating LTL using the proposed method?**
Selecting the appropriate specification template (4 of 10)
Replacing the states in the mapping template with parameters in the original property specification (1 of 10)
3. **Compared to the existing method, what are the advantages of the proposed method?**
The proposed method is faster and easier (7 of 10)
The proposed method does not require the understanding of the pattern system (4 of 10)
4. **Please tell us if you have any idea of improving our method of creating LTLs**
The specification template should be easier to understand and select (4 of 10)
There should be a way to check if the generated LTL is correct (3 of 10)

7 Threats to Validity

Participants with different levels of understanding of LTLs and pattern systems might be an internal threat to validity. We did not evaluate participants' knowledge of LTLs or pattern systems. Instead, participants were randomly selected. Additionally, participant A and B failed to record the accurate required times, which might also affect the results of RQ1. Although we tried to keep corresponding questions (e.g., Q1 and Q5) at the same level of difficulty, they were not identical to each other.

As we mentioned before, our original research and experiment were implemented in Japanese. Therefore, the results may not be the same when property specifications are written in different languages, which is an external threat to validity.

8 Related Work

Some studies were proposed to help non-experts use specification patterns in formal methods. Nelken and Francez [14] present a method for automatically translating natural language specification into temporal logic. Flake et al. [15] struct English

sentences for Clocked CTL [15]. Konrad and Cheng [9] develop structured English grammars and tool support for their property specification patterns. However, existing research lack quantitative evaluation. Our study provides a quantitative evaluation with structured Japanese grammars.

Numerous approaches of temporal logic visualization have been studied. Hassine et al. [5] present a graphical specification pattern catalog based on the UCM notation [16], which covers all qualitative specification patterns introduced by Dwyer et al. [7]. Giannakopoulou and Magee [17] develop tool support (LTSA) for model checking fluent-based LTL properties. Feja and Fötsch [18] propose a visual notation for logical rules at the level of processes and workflows. Although existing methods have been proved useful, they require specific tool support. Our proposal focuses on simple graphics which could be easily implemented on any platform.

9 Conclusion

Temporal logics are widely used in software verification such as model checks. However, unfamiliarity with property specifications and notations of temporal logics makes it difficult to create temporal logics like LTL. Dwyer et al. proposed a pattern system for nonexpert practitioners to create temporal logics correctly. Nevertheless, their approach is time consuming and error-prone, especially when the specifications become complicated.

In this research, we propose a method to create LTLs based on property specification patterns. With our approach, practitioners refine the original property specifications to a given format. Then our method generates a real LTL automatically based on the refined property specifications. An experiment confirmed that our approach can improve both the time cost and the accuracy compared to the existing method. We also introduced PSG after the experiment. Practitioners can use PSGs to check if the generated LTLs are correct.

The questionnaire revealed several directions to improve our approach. To enhance practitioners' understanding, we plan to devise property specification templates that more easily distinguish between each pattern and scope. We also are considering building a pattern prediction system to help practitioners select the appropriate property specification templates. Finally, we are developing an experiment to assess the usefulness of PSG.

Acknowledgements This research is sponsored by the DENSO CORPORATION. We would like to thank S. Takahashi, A. Hata, S. Tanaka, K. Kanamori, and N. Kishimoto for their helpful input. We also appreciate other anonymous participants in our experiment for their time and help.

References

1. McMillan, K.L.: Symbolic Model Checking, pp. 25–60. Symbolic Model Checking. Springer, Boston, MA (1993)
2. Manna, Z., Pnueli, A.: The Temporal Logic of Reactive and Concurrent Systems: Specification. Springer Science & Business Media (2012)
3. Clarke, E.M., Emerson, E.A., Sistla, A.P.: Automatic verification of finite-state c oncurrent systems using temporal logic specifications. ACM Trans. Program. Lang. Syst. (TOPLAS) **8**(2), 244–263 (1986)
4. De Nicola, R., Fantechi, A., Gnesi, S., Ristori, G.: An action-based framework for veryfying logical and behavioural properties of concurrent systems. Comput. Netw. ISDN Syst. **25**(7), 761–778 (1993)
5. Hassine, J., Rilling, J., Dssouli, R.: Use Case Maps as a property specification language. Softw. & Syst. Model., pp. 205–220 (2009)
6. Dwyer, M.B., Avrunin, G.S., Corbett, J.C.: Property specification patterns for finite-state verification. In: Proceedings of the Second Workshop on Formal Methods in Software Practice. ACM (1998)
7. Dwyer, M.B., Avrunin, G.S., Corbett J.C.: Patterns in property specifications for finite-state verification. In: Proceedings of the 21st International Conference on Software Engineering. ACM (1999)
8. Gruhn, V., Laue, R.: Specification patterns for time-related properties. Temporal representation and reasoning. In: TIME 2005, 12th International Symposium on. IEEE (2005)
9. Konrad, S., Cheng, B.H.C.: Facilitating the construction of specification pattern-based properties. Requirements Engineering, 2005. In: Proceedings. Conference on 13th IEEE International. IEEE (2005)
10. Maezawa, Y., Washizaki, H., Tanabe, Y., Honiden, S.: Automated verification of pattern-based interaction invariants in Ajax applications. In: Proceedings of 28th IEEE/ACM International Conference on Automated Software Engineering (ASE2013), pp. 158–168 (2013)
11. Kubo, A., Washizaki, H., Fukazawa, Y.: Automatic extraction and verification of page transitions in a web application. In: 14th Asia-Pacific Software Engineering Conference (APSEC'07), pp. 350–357 (2007)
12. UMLet project: https://www.umlet.com/
13. OSEK/VDX Operating System Specification 2.2.3, http://www.irisa.fr/alf/downloads/puaut/TPNXT/images/os223.pdf (2005)
14. Nelken, R., Francez, N.: Automatic translation of natural language system specifications into temporal logic. In: International Conference on Computer Aided Verification. Springer, Berlin, Heidelberg (1996)
15. Flake, S., Müller, W., Ruf J.: Structured English for Model Checking Specification. MBMV (2000)
16. Buhr, R.J.A., Casselman, R.S.: Use Case Maps for Object-Oriented Systems. Prentice-Hall, Inc. (1995)
17. Giannakopoulou, D., Magee, J.: Fluent model checking for event-based systems. ACM SIGSOFT Software Engineering Notes, vol. 28, no. 5. ACM (2003)
18. Feja, A., Fötsch, S.: "Model checking with graphical validation rules." Engineering of Computer Based Systems, 2008. ECBS 2008. In: 15th Annual IEEE International Conference and Workshop on the. IEEE (2008)

Rule-Based BCG Matrix for Product Portfolio Analysis

Chih-Chung Chiu and Kuo-Sui Lin

Abstract Product portfolio analysis is used for analyzing company products' strategic market position, in order to decide which products should receive more or less investment. The Boston Consulting Group (BCG) growth-share matrix is the best known approach for product portfolio analysis. It is a strategic tool for identifying products' strategic market positions and formulating resources allocation strategies. However, traditional BCG matrix is a static historic analysis, which different time frames are not considered. Thus, the main purpose of this study was to propose a new BCG method, which is a dynamic trend analysis method and different time frames are considered for product portfolio analysis. A numerical case study was conducted to demonstrate that the proposed BCG method has achieved the objectives of the study for company products' portfolio analysis. First, the rule-based BCG method has contributed a new vague set based data collection method. Second, the rule-based BCG method has contributed a new application for identifying company products' strategic market positions and resources allocation strategy formulation.

Keywords BCG growth-share matrix · Product portfolio analysis · Rule-based classification · Vague set theory

1 Introduction

In the company, the problem of which product opportunity and how much resource to commit is an ever going headache for those who allocate the scarce resources. Product portfolio analysis is an increasingly popular approach for strategic planning within multiproduct corporations. A product portfolio is the collection of all the products (or services) offered by a company. To maintain sustainable competitive advantage,

C.-C. Chiu (✉) · K.-S. Lin
Department of Information Management, Aletheia University, 25103 New Taipei City, Taiwan (R.O.C.)
e-mail: au4229@mail.au.edu.tw

K.-S. Lin
e-mail: au4234@mail.au.edu.tw

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_2

a company should have a portfolio of products with different market size growth rate and market share growth rate. Product portfolio analysis is used to assist in analyzing an existing product portfolio to decide which products should receive more or less investment, and adding new products to the portfolio or deciding which products and businesses should be eliminated. The Boston Consulting Group (BCG) growth-share model is the best known [1] approach to portfolio analysis, for prioritizing them and allocating resources.

However, BCG matrix is only a snapshot of the current marketing position and it has little or no predictive value for trend analysis. Besides, market size growth rate is a collective subjective measure of a specific company product's market's opportunity. Relative market share growth rate is a collective subjective measure of a specific company product's capability to drive competitive advantage. Thus the main purpose of this study was to propose a new BCG product portfolio analysis method, which is a collective and dynamic method to the product portfolio analysis panel for identifying company products' strategic market positions and formulating resources allocation strategies under vague and uncertain environment. Rather than BCG matrix's static historic analysis, it provides a collective forecasting method to the portfolio analysis panel for forecasting company products' expected market size growth rates and expected relative market share growth rates. It also provides recommendation strategies, depending on which strategy quadrants the products fall, to help company make resource allocation decisions.

Thus, he novel contributions of this paper are: (1) to propose a vague set based data collection method that solicits vague values of a specific company product's market size growth rate and relative market share growth rate under vague and uncertain environment; (2) to propose a dynamic BCG method to the product portfolio analysis panel for identifying the company product's strategic market position and formulating resources allocation strategy.

2 Theoretical Background

2.1 BCG Product Portfolio Analysis Model

The BCG matrix (BCG Product portfolio Analysis model, BCG product portfolio matrix or BCG Growth Share Matrix) is a product portfolio analysis model created by Henderson [2] for the Boston Consulting Group (https://www.bcg.com) to help enterprise make resource allocation decisions related to their SBUs/products portfolio strategies. BCG Growth-Share Matrix is plotted on a two-dimensional four celled grid (2×2 *matrix*). Using the BCG grid, a company classifies all its SBUs/Products according to two dimensions: (1) on the horizontal axis is the *Relative Market Share* (*RMS*)—this serves as a measure of product strength (competitive advantage) in the market. The market share of the product in the market is defined as compared to

its competitors and overall product/category; (2) on the vertical axis is the *Market Growth Rate* (*MGR*)—this provides a measure of market attractiveness.

A company product's *Market Growth Rate* provides an external measure of market attractiveness and a company product's *Relative Market Share* serves as an internal measure of company strength in the market. A company product's market growth rate is a measure of the size of the market opportunity a company operates in is growing or shrinking. It is defined as the rise in sales or market size within a given customer base over a specific period of time. A company product's relative market share is a marketing metric used to define percentage of the company's product market share within a given market or industry. Relative market share allows a company (and their investors) to see how the company's product is successful and its position in the market. A company with a declining relative market share might be viewed as undesirable from an investment standpoint. By contrast, a company with a growing market share is indicative of a company's competitive advantage. By dividing the matrix into four cells (quadrants), four categories of strategic market position can be classified, shown as: star, problem child, cash cow and dog (Table 1).

As shown in Fig. 1, it applies two inputs, *Market Growth Rate* (*MGR*) and *Relative Market Share* (RMS), to a portfolio of products or strategic business units, and then draws conclusions about how resources (e.g. talent, investment) should be allocated across the portfolio. The *Market Growth Rate* is shown on the vertical (y) axis and the *Relative Market Share* axis is shown on the horizontal (x) axis. The range is set somewhat arbitrarily. A product's *MGR* and *RMS* lead to a classification into one of four categories of product market positions.

The four basic resource allocation choices that help to implement the four categories of product positions are: (a) *Build* Strategy: Make further investments to Build Market Share (for example, to maintain Star status, or turn a *QUESTION MARK* into a *STAR*); (b) *Hold* Strategy: Maintain the status quo (do nothing) to preserve market share; (c) *Harvest* Strategy: Reduce the investment (enjoy positive cash flow and maximize profits from a *STAR* or *CASH COW*); (d) *Divest* Strategy: To sell or liquidate the business because resources can be better used elsewhere. For example, get rid of the *DOGs*, and use the capital to invest in *STAR*s and some *QUESTION MARK*s.

2.2 Vague Set Theory

The intuitionistic fuzzy set (IFS) on a universe X was introduced by Atanassov [3, 4] as a generalization of fuzzy set introduced by Zadeh in 1965 [5]. An intuitionistic fuzzy set $\tilde{A}$ in X is an object having the form $A = \{<x, \mu_A(x), \nu_A(x) > | x \in X\}$, which is characterized by a membership function (true membership function) μ_A and a non-membership function (false membership function) ν_A, where the function μ_A: $X \rightarrow [0, 1]$ and $\nu_A(x)$:$X \rightarrow [0, 1]$ define the degree of membership and degree of non-membership of the element $x \in X$ to the set A, respectively. Later, Gau and Buehrer [6] proposed the concept of vague set (*VS*), where the grade of membership is bounded

Table 1 Presentations of the BCG growth share matrix

Strategic product position	Market growth rate (MGR)	Relative market share (RMS)	Prescription
STAR product position	Good	High	*STAR* is a SBU/product with a high market share in a fast-growing industry. The main focus of *STAR* businesses is to protect their market shares and thus get a bigger portion of the market growth than competitors
CASH-COW product position	Bad	High	*CASH COW* is a SBU/product with high market share in a slow-growing or shrinking industry. The unit is characterized by high profit and cash generation. Typically, it generates excess cash above the amount of cash needed to maintain the business. *CASH COW* is to be milked continuously with as little investment as possible, since such an investment would be wasted in an industry with low growth
PROBLEM- CHILD product position	Good	Low	*PROBLEM CHILD* is a SBU/product with a low market shares in a fast-growing industry. Because the market is growing rapidly and thus consumes large amounts of cash, but because the unit has low market share they do not generate much cash. The result is that the unit does not generate much cash and has high cash needs
DOG product position	Bad	Low	*DOG* is a SBU/product with a high market share in a mature, slow-growing industry or depressed, shrinking industry. Often generate poor profits and cash needs are frequently higher than the cash that is generated. To improve the overall performance, firms should minimize the proportion of their assets, harvesting by cutting costs and maximizing cash flow by divestment or liquidation. Even worse, the unit should be sold off

	RELATIVE MARKET SHARE: High ≥100%	RELATIVE MARKET SHARE: Low 50% 0%
MARKET GROWTH RATE: Good +20% 0%	*STAR*	*PROBLEM CHILD*
MARKET GROWTH RATE: Bad −20%	*CASH COW*	*DOG*

Fig. 1 BCG product portfolio matrix. *Source* Adapted from Boston Consulting Group [10]

to a subinterval $[t_A(x), 1 - f_A(x)]$ of [0, 1]. Burillo and Bustince [7] proved that the notion of vague sets coincides with that of intuitionistic fuzzy sets. Relevant definition and operations of vague sets introduced in [6, 8] are briefly reviewed as follows.

Definition: Vague Set

A vague set A in the universe of discourse X is characterized by a truth membership function, $t_A: X \rightarrow [0, 1]$, and a false membership function, $f_A: X \rightarrow [0, 1]$, where $t_A(x)$ is a lower bound of the grade of membership of x derived from the "evidence for x", and $f_A(x)$ is a lower bound on the negation of x derived from the "evidence against x", and $0 \leq t_A(x) + f_A(x) \leq 1$. The grade of membership of x in the vague set is bounded to a subinterval $[t_A(x), 1 - f_A(x)]$ of [0, 1].

The interval $[t_A(x), 1 - f_A(x)]$ is called the vague value of x in A and can be denoted by $V_A(x)$. The vague value $[t_A(x), 1 - f_A(x)]$ indicates that the exact grade of membership $\mu_A(x)$ of x may be unknown, but is bounded by $t_A(x) \leq \mu_A(x) \leq 1 - f_A(x)$, where $t_A(x) + f_A(x) \leq 1$. This interval can be interpreted as an extension to the fuzzy membership function. The precision of uncertainty about x is characterized by the difference between $1 - f_A(x)$ and $t_A(x)$, i.e., $1 - f_A(x) - t_A(x)$. The value of $\pi_A(x) = 1 - f_A(x) - t_A(x)$ *expresses* a hesitation degree of *whether x belongs* to A or *not*. If this value is small, the knowledge about x is relatively precise; if this value is large the knowledge about x is little. If $t_A(x)$ is equal to $1 - f_A(x)$, there is no hesitation and the vague set theory reduces to the fuzzy set theory. If the universe of discourse X is a finite set, then a vague set A of the universe of discourse X can be written as $A = \{(x, [t_A(x), 1 - f_A(x)]) | x \in X\}$. Let $f^*_A(x) = 1 - f_A(x)$, the vague value $V_A(x) = [t_A(x), 1 - f_A(x)] = [t_A(x), f^*_A(x)]$. Then, the unknown part $\pi_A(x)$ of the vague value $V_A(x)$ is defined as follows: $\pi_A(x) = 1 - f_A(x) - t_A(x) = f^*_A(x) - t_A(x)$.

3 Proposed New Method for Forecasting Market Size Growth Rate and Market Share Growth Rate

BCG analysis is mainly used for a multi-product company to help in deciding which product in the product portfolio needs building (increasing investment), which needs harvesting (harvesting money), which needs divesting (reducing investment) and which needs to be watched. The BCG matrix considers two variables, namely, market size growth rate and market share growth rate. In this section, a vague set based method for forecasting market sales size growth rate and market share growth rate is briefly reviewed. The procedure is presented as follows:

Step 1: *Soliciting vague values of company products' market growth rates and products' Shares*

The vague membership values of market growth rate in the vague grade sheet can be solicited in this step. A *company's* financial forecasting panel can employ the new polling method proposed by the author (Lin and Chiu, [9] to solicit vague values $V(Q_{ij})$ of the j-th grade for the i-th product Q_i $(i = 1 \ldots m)$. The vague membership values $V(Q_{ij})$ $(i = 1 \ldots m; j = 1 \ldots n)$ are filled in the vague grade sheet as shown in Table 2, where $V(Q_{ij})$ denotes the solicited growth rate of the j-th level grade G_j for the i-th product Q_i. The solicited vague values in the forecasting vague grade sheet can be regarded as a set of vague values, where each element in the universe of discourse belonging to a vague set is represented by a vague value in [0, 1].

Step 2: *Transforming numerical scores of the Solicited vague values*

In order to compute individual market growth rate of *company's* specific product Q_i, the solicited vague membership values $V(Q_{ij})(i = 1, \ldots, m; j = 1, \ldots, n)$ in the vague evaluation sheet must be transformed into numerical scores in this step. The new score function proposed by the author [9] can be used to transform the vague values $V(Q_{ij})$ into numerical scores. Thus, the numerical score of the vague value can be calculated by the following transforming score function:

$$S_L\left(V\left(Q_{ij}\right)\right) = t_A\left(x_{ij}\right)/2 + \left(1 - f_A\left(x_{ij}\right)\right)/2 = \left(t_A\left(x_{ij}\right) + f_A^*\left(x_{ij}\right)\right)/2 \quad (1)$$

Table 2 Vague Grade Sheet

Product no.	Vague value					
	G_1	G_2	…	G_j	…	G_n
Q_1	$V(Q_{11})$	$V(Q_{12})$	…	$V(Q_{1j})$	…	$V(Q_{1n})$
Q_2	$V(Q_{21})$	$V(Q_{22})$	…	$V(Q_{2j})$	…	$V(Q_{2n})$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Q_3	$V(Q_{i1})$	$V(Q_{i2})$	…	$V(Q_{ij})$	…	$V(Q_{in})$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Q_m	$V(Q_{m1})$	$V(Q_{m2})$	…	$V(Q_{mj})$	…	$V(Q_{mn})$

Step 3: *Deriving the ranges of level grades into numerical scores*

A set of level grades are used to forecast the *company's* growth rate forecasting regarding a product portfolio Q, where $Q \in \{Q_1, Q_2,\ldots,Q_i,\ldots,Q_m\}$. Assume that the set of level grade assigned to the product Q_i is G, where $G_j \in \{G_1, G_2,\ldots,G_j,\ldots,G_n\}$ and $0\% \leq m_{1j} \leq E(G_j) \leq m_{2j} \leq 100\%$, then the expected growth rates against the j-th level grade G_j are evaluated for each product Q_i as follows:

$$E\left(G_j\right) = (1 - \lambda) \times m_{1j} + \lambda \times m_{2j}, \tag{2}$$

where $\lambda \in [0, 1]$ denotes the optimism index determined by the evaluator, and $E(\mathrm{G}_j)$ is the expected growth rate against the j-th level grade G_j. If $0 \leq \lambda \leq 0.5$, the evaluator is a pessimistic evaluator. If $\lambda = 0.5$, the evaluator is a normal evaluator. If $0.5 \leq \lambda \leq 1.0$, the evaluator is an optimistic evaluator.

Step 4: *Computing the company product's expected market size growth rate*

A company product's expected market size growth rate $EMGR(Q_i)$ is evaluated for each product Q_i as follows:

$$\begin{aligned} EMGR(Q_i) &= \sum_{j=1}^{n}\left[S_L\left(V\left(Q_{ij}\right)\right) \times E\left(G_j\right)\right] \sum_{j=1}^{n} S_L\left(V\left(Q_{ij}\right)\right) \\ &= [S_L(V(Q_{i1})) \times E(G_1) + S_L(V(Q_{i2})) \\ &\quad \times E(G_2) + \cdots + S_L\left(V\left(Q_{ij}\right)\right) \times E\left(G_j\right) \\ &\quad + \cdots + S_L(V(Q_{\mathrm{in}})) \times E(G_n)] \\ &\quad /[S_L(V(Q_{i1})) + S_L(V(Q_{i2})) + \cdots + S_L(V(Q_{in}))] \end{aligned} \tag{3}$$

where $S_L(V(O_{ij}))$ denotes the transformed market growth rate score against the j-th level grade G_j for the i-th product Q_i, and $E(\mathrm{G}_j)$ is the expected market growth rate against the j-th level grade G_j.

Step 5: *Repeating Step* 1 *through Step* 3 *to compute the company product's expected relative market share ERMS*(Q_i).

4 Proposed New BCG Matrix for Product Portfolio Analysis

BCG matrix helps companies figure out which areas of their products deserve more resources and investment. Based on the above new forecasting method for expected market size growth rate and expected market share growth rate, a new BCG matrix for product portfolio analysis is presented in this section, described as follows:

Stage 1: *Measurement Stage*

In the BCG product portfolio analysis matrix, a company classifies its different products on a two dimensional growth-share matrix. In the BCG matrix, vertical axis represents Expected Market Growth Rate *EMGR*(Qi) and horizontal axis represents Expected Relative Market Share *ERMS*(Qi). Company product's Expected Market Growth Rate provides an external measure of market attractiveness and company product's Expected Relative Market Share serves as an internal measure of company strength in the market.

Expected market size growth rate of a product is the forecasted total market size growth rate compared to its present total market size: Expected market size growth rate (for subsequent period t) = {[total market sales size (for subsequent period t)—total market sales size (for current period $t-1$)]/Total market sales size (for current period $t-1$)}:

$$EMGR_t = ((TMS_t - TMS_{t-1})/TMS_{t-1}) \times 100\%. \tag{4}$$

Expected relative market share (*ERMS*) can be defined as the percentage of a market's total sales, which is earned by a particular company product over a specified time period. Expected relative market share is calculated by taking a particular company product's sales over the period and dividing it by the total market sales over the same period: Expected Relative Market Share (for period t) = company product's sales (for period t)/total market sales of the product (for period t):

$$ERMS_t = (CPS_t/TMS_t) \times 100\%. \tag{5}$$

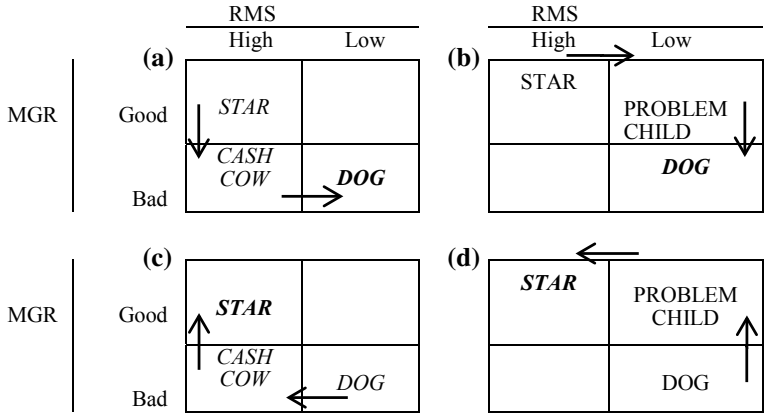
Stage 2: *Classification Stage*

Using the BCG Matrix, a company can classify its different Strategy Quadrants of products as follows:

(a) *Cash Cow* Product Position: High market share business in low market growth market (most profitable)—Taking Milking strategy. (b) *Star* Product Position: High market share business in high-growth market (high competition)—Taking four growth strategies: Product development strategy, Market development strategy, Market penetration strategy and Diversification strategy. (c) *Problem Child* (*Question Marks or Wild Cats*) Product Position: Low market share business in high-growth market (uncertainty)—Taking Selective investment strategy. (d) *Dog* Product Position: Low market share business in low growth rate market (less profitable or may even be negative profitability)—Taking Exit strategy.

However, traditional BCG matrix is a static analysis, which different time frames are not considered. In the real world business, static BCG matrix should be complemented by dynamic analysis in order to incorporate the progress route of company's strategic product positions over different periods of time.

Figure 2 depicts the dynamic BCG matrix in terms of progress route of company's product positions in the Matrix. A company product's *Market Growth Rates* and



Note: (a)(b): Pessimistic Progress Route (from *STAR* to *DOG*);
(c)(d): Optimistic Progress Route y (from *DOG* to *STAR*)

Fig. 2 Company product's progress routes in the dynamic BCG matrix

Relative Market Shares can either increase or decrease in different time frames. This means the product's associated market positions may change and produce a progress route. In order to track the movement of a product over time in the BCG matrix, a negative progress route and a positive progress route are proposed and defined in this study. The negative progress route of the dynamic BCG matrix is defined as a pessimistic progress route of a product in BCG matrix. If any negative progress route happens, it means a product moves from a *STAR* to a *Problem Child*, and finally to a *Dog* because of its decreasing forecasted *Relative Market Share* and/or decreasing forecasted *Market Growth Rate*. Unlike the negative progress route of BCG matrix, the positive progress route of BCG matrix is defined as an optimistic progress route of a product in BCG matrix. If any positive progress route happens, it means a product moves from a *Dog* to a *Problem Child* (or *CASH COW*), and finally to a *Star* because of its increasing market share and/or increasing *Market Growth Rate* situation.

The trend of a product's progress routes in the dynamic BCG matrix (Fig. 2) provides expert knowledge involved in the BCG product portfolio analysis. The derived knowledge will lead the analysis panel to locate company products' strategic product positions and to formulate resource allocation strategies. Based on the trend of a product's progress routes in the dynamic BCG matrix, a set of classification rules is formulated in this study to identify strategic market position of the product. Rule-based classification refers to any classification scheme that makes use of IF-THEN rules for class prediction. IF-THEN rules are expressed in the syntax as: IF <antecedent>, THEN <consequent>. As shown in Table 3, the formulated classification rules can be obtained to describe the domain knowledge involved in the new product portfolio analysis.

Stage 3: *Recommendation Stage*

Table 3 Rule-based knowledge for product portfolio analysis

Rule no.	IF-THEN statement
Rule R1	IF Market Growth Rate is "Good+" and Market Share is "High+",THEN Strategy Quadrant is "*STAR*" and the Recommended Strategy is "*Build*"
Rule R2	IF Market Growth Rate is "Good−" and Market Share is "High−",THEN Strategy Quadrant is "*STAR*" and the Recommended Strategy is "*Hold*"
Rule R3	IF Market Growth Rate is "Good+" and Market Share is "High−",THEN Strategy Quadrant is "*STAR*" and the Recommended Strategy is "*Build* or *Hold*"
Rule R4	IF Market Growth Rate is "Good +" and Market Share is "Low+", THEN Strategy Quadrant is "*PROBLEM CHILD*" and the Recommended Strategy is "*Build* or *Hold*"
Rule R5	IF Market Growth Rate is "Good−" and Market Share is "Low−", THEN Strategy Quadrant is "*PROBLEM CHILD*" and the Recommended Strategy is "*Divest*"
Rule R6	IF Market Growth Rate is "Good +" and Market Share is "Low–", THEN Strategy Quadrant is "*PROBLEM CHILD*" and the Recommended Strategy is "*Hold* or *Divest*"
Rule R7	IF Market Growth Rate is "Bad+" and Market Share is "High+", THEN Strategy Quadrant is "*CASH COW*" and the Recommended Strategy is "*Build*, *Hold* or *Harvest*"
Rule R8	IF Market Growth Rate is "Bad −" and Market Share is "High−", THEN Strategy Quadrant is "*CASH COW*" and the Recommended Strategy is "*Harvest*"
Rule R9	IF Market Growth Rate is "Bad +" and Market Share is "High−", THEN Strategy Quadrant is "*CASH COW*" and the Recommended Strategy is "*Hold* or *Harvest*"
Rule R10	IF Market Growth Rate is "Bad +" and Market Share is "Low+", THEN Strategy Quadrant is "*DOG*" and the Recommended Strategy is "*Hold* or Build"
Rule R11	IF Market Growth Rate is "Bad–" and Market Share is "Low−", THEN Strategy Quadrant is "*DOG*" and the Recommended Strategy is "*Divest*"
Rule R12	IF Market Growth Rate is "Bad +" and Market Share is "Low −", THEN Strategy Quadrant is "*DOG*" and the Recommended Strategy is "*Hold* or *Divest*"

After the classification stage, the classified rule recommends the matching strategy, depending on which strategy quadrant the product fall and whether it is in a pessimistic progress or an optimistic progress. As shown in Table 4, a product with positive market size growth rate and positive relative market share, the product is classified as a *Star* product. Thus, a *Build* or *Hold* resource allocation strategy is recommended.

5 Numerical Case Study

A manufacturer M was taken as a case study company. The company M manufactures a group of related products (referred to as product portfolio or product line Q) under a single brand. The company M planned to make a product portfolio analysis of company's product portfolio Q, $Q \in \{Q_1, Q_2, Q_3, Q_4\}$. The proposed method of

Table 4 Strategy quadrant and recommended strategy

Rule no.	Market growth rate	Relative market share	Strategic product position	Resource allocation strategy
Rule R1	G+	H+	*Star*	*Build or Hold*
Rule R2	G–	H–	*Star*	*Hold*
Rule R3	G +	H–	*Star*	*Build or Hold*
Rule R4	G+[a]	L+	*Problem Child*	*Build, Hold or Divest*
Rule R5	G –	L–[b]	*Problem Child*	*Divest*
Rule R6	G +	L–	*Problem Child*	*Hold or Divest*
Rule R7	B+	H+	*Cash Cow*	*Build, Hold or Harvest*
Rule R8	B–	H–	*Cash Cow*	*Harvest*
Rule R9	B+	H–	*Cash Cow*	*Hold or Harvest*
Rule R10	B+	L+	*Dog*	*Hold or Build*
Rule R11	B–	L–	*Dog*	*Divest*
Rule R12	B+	L–	*Dog*	*Hold or Divest*

Note: [a]G+: Good and increasing market growth rate
[b]B−: Bad and decreasing market growth rate

this study was employed for company's portfolio analysis. As shown in Table 5, the seven level grades used for evaluating company's product *Expected Market Growth Rates* are: $G_{MGR} = \{G_1, G_2, G_3, G_4, G_5, G_6, G_7\} = \{$*Very Very Bad* (*VVB*), *Very Bad* (*VB*), *Bad* (*B*), *Moderate* (*M*), *Good* (*G*), *Very Good* (*VG*), *Very Very Good* (*VVG*)$\}$. The optimism index λ determined by the evaluator is 0.50 (i.e., $\lambda = 0.50$). Based on Eq. (2), the expected growth rate $E(G_j)$ of the assigned level grade G_j for each product is calculated as follows:

$$E(G_7) = (1 - 0.50) \times 50\% + 0.50 \times 100\% = 75\%,$$

Table 5 Linguistic level for achievement of company product's market growth rate

Grade	Linguistic level of achievement of MGR	Range of growth rate (%)	Expected growth rate $E(G_j)$ (%)
G_7	*Very Very Good (VVG)*	50–100	75
G_6	*Very Good (VG)*	25–50	37.5
G_5	*Good (G)*	0–25	12.5
G_4	*Moderate (M)*	0	0
G_3	*Bad (B)*	−25–0	−12.5
G_2	*Very Bad (VB)*	−26–50	−37.5
G_1	*Very Very Bad (VVB)*	−50–100	−75

$$E(G_6) = (1 - 0.50) \times 25\% + 0.50 \times 50\% = 37.5\%,$$
$$E(G_5) = (1 - 0.50) \times 0\% + 0.50 \times 25\% = 12.5\%,$$
$$E(G_4) = (1 - 0.50) \times 0\% + 0.50 \times 0\% = 0\%,$$
$$E(G_3) = (1 - 0.50) \times 0\% + 0.50 \times (-25\%) = -12.5\%,$$
$$E(G_2) = (1 - 0.50) \times (-25\%) + 0.50 \times (-50\%) = -37.5\%,$$
$$E(G_1) = (1 - 0.50) \times (-50\%) + 0.50 \times (-100\%) = -75\%.$$

A product portfolio analysis panel used a vague value sheet as shown in Table 6 to forecast company products' *Expected Market Growth Rates*.

It indicated that the forecasted vague values of the first product Q_1 are: [0.4, 0.5] for *Good* and [0.8, 0.9] for *Very Good*. Using Eq. (1), the solicited level-grade vague membership value of the product $V(Q_{ij})$ can be transformed into level-grade numerical score $S(V(Q_{ij}))$. The transformed level-grade numerical scores for products Q_1, Q_2 and Q_3 are summarized in Table 7.

From Table 7, the forecasted vague values of the products Q_1, Q_2, Q_3 and Q_4 can be represented by vague sets shown as follows:

$Q_1 = \{(VVB, [0, 0]), (VB, [0, 0]), (B, [0, 0]), (M, [0, 0]), (G, [0.4, 0.5]),$
$(VG, [0.8, 0.9]), (VVG, [0, 0])\}.$

$Q_2 = \{(VVB, [0, 0]), (VB, [0, 0]), (B, [0.4, 0.5]), (M, [1, 1]), (G, [1, 1]),$
$(G, [0.6, 0.7]), (VG, [0.4, 0.5]), (VVG, [0, 0])\}.$

$Q_3 = \{(VVB, [0, 0]), (VB, [0.8, 0.9]), (B, [0.5, 0.6]), (M, [0.2, 0.3]), (G, [0, 0]),$
$(VG, [0, 0]), VGG, [0, 0])\}.$

Table 6 Vague graded evaluation sheet for evaluating product's market growth rate

Product no.	Graded vague membership value for product's MGR						
	VVB	*VB*	*B*	*M*	*G*	*VG*	*VVG*
Q_1	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0.4, 0.5]	[0.8, 0.9]	[0, 0]
Q_2	[0, 0]	[0, 0]	[0.4, 0.5]	[1, 1]	[0.6, 07]	[0.4, 0.5]	[0, 0]
Q_3	[0, 0]	[0.8, 0.9]	[0.5, 0.6]	[0.2, 0.3]	[0, 0]	[0, 0]	[0, 0]

Table 7 Transformed vague graded evaluation sheet for evaluating product's market growth rate

Product no.	Graded score for product's MGR						
	VVB	VB	B	M	G	VG	VVG
Q_1	0	0	0	0	0.45	0.85	0
Q_2	0	0	0.45	1	0.65	0.45	0
Q_3	0	0.45	0.55	0.25	0	0	0

Using Eq. (1), the forecasted vague values of the products Q_1, Q_2, Q_3 and Q_4 can be transformed into numerical scores. Using Eq. (3), the *Expected Market Growth Rates* for the first product Q_1 can be calculated as:

$$\begin{aligned} EMGR(Q_1) = [&S_L(V(Q_{11})) \times E(VVB,) + S_L(V(Q_{12})) \times E(VB) \\ &+ S_L(V(Q_{13})) \times E(B) + S_L(V(Q_{14})) \times E(M) \\ &+ S_L(V(Q_{15})) \times E(G) + S_L(V(Q_{16})) \\ &\times E(VG) + S_L(V(Q_{17}))][S_L(V(O_{11})) \\ &+ S_L(V(O_{12})) + S_L(V(O_{13})) + S_L(V(Q_{14})) \\ &+ S_L(V(Q_{15})) + S_L(V(Q_{16})) + S_L(V(Q_{17}))] \\ = &(0.45 * 12.5\% + 0.85 * 37.5\%))/(0.45 + 0.85) = 28.85\%. \end{aligned}$$

Similarly, the *Expected Market Growth Rates* for the product Q_2 and Q_3 can be calculated as:

$$\begin{aligned} EMGR(Q_2) = &\big(0.45^*(-12.5\%) + 1 * 0\% + 0.65 * (12.5\%) + 0.45 * (37.5\%)\big) \\ &/(0.45 + 1 + 0.65 + 0.45) = 7.6\%, \end{aligned}$$

$$\begin{aligned} EMGR(Q_3) = &(0.85 * (-37.5\%) + 0.55 * (-12.5\%) + 0.25 * 0\%) \\ &/(0.85 + 0.55 + 0.25) = -23.48\%. \end{aligned}$$

Repeating step 1 through step 3, the product portfolio analysis panel used the vague value forecasting sheet to forecast company products' *Expected Relative Market Share*s, as shown in Table 8.

The range of *Expected Relative Market Share* and their associated original grade and level of achievement is shown in Table 9.

As shown in Table 9, the seven linguistic grades used for evaluating company's product *Relative Market Share*s are: $G_{RMS} = \{G_1, G_2, G_3, G_4, G_5, G_6, G_7\} = \{$*Very Very Low*(*VVL*), *Very Low*(*VL*), *Low*(*L*), *Moderate*(*M*), *Good*(*G*), *Very Good*(*VG*), *Very Very Good* (*G*)(*VVG*)$\}$.

It indicated that the forecasted vague values of the *Expected Relative Market Shares* for the company's first product Q_1 are: [0.8, 0.9] for *Low*, [0.6, 0.8] for

Table 8 Vague graded evaluation sheet for evaluating product's relative market share

Product no.	Graded vague membership value for RMS						
	VVL	VL	L	M	H	VH	VVH
Q_1	[0, 0]	[0, 0]	[0.8, 0.9]	[0.6, 0.8]	[0.5, 0.7]	[0, 0]	[0, 0]
Q_2	[0, 0]	[0, 0]	[0, 0]	[04, 0.5]	[0.6, 0.7]	[1, 1]	[0, 0]
Q_3	[0, 0]	[0, 0]	[0.4, 0.5]	[0.8, 0.9]	[0.6, 07]	[0.4, 0.5]	[0, 0]

Table 9 Linguistic level for achievement of company product's relative market share

Grade	Linguistic level of achievement of RMS	Range of market share (%)	Expected market share E(Gj) (%)
G_7	*Very Very High (VVH)*	100	100
G_6	*Very Good(VH)*	80–100	90
G_5	*Good (H)*	60–80	70
G_4	*Moderate(M)*	40–60	50
G_3	*Low (L)*	20–40	30
G_2	*Very Low (VL)*	0–20	10
G_1	*Very Very Low(VVL)*	0	0

Moderate and [0.5, 0.7] for *High*. From Table 8, the forecasted vague values of the products Q_1, Q_2 and Q_3 can be represented by vague sets shown as follows:

$Q_1 = \{(VVL, [0, 0]), (VL, [0, 0]), (L, [08, 0.9]), (M, [[0.6, 0.8]), (H, [0.5, 0.7]), (VH, [0, 0]), (VVH, [0, 0])\}$.

$Q_2 = \{(VVL, [0, 0]), (VL, [0, 0]), (L, [0, 0]), (M, [0.4, 0.5]) (H, [0.6, 0.7]), (VH_2[1, 1]), (VVH, [0, 0])\}$.

$Q_3 = \{(VVL, [0, 0]), (VL, [0, 0]), (L, [0.4, 0.5]), (M, [0.8, 0.9]), (H, [0.6, 0.7]), (VH, [0.4, 0.5]), (VVH, [0, 0])\}$.

Using Eq. (3), the forecasted vague values of the *Expected Relative Market Share* for the products Q_1 can be transformed into a numerical score:

$$\begin{aligned} ERMS(Q_1) = {} & S_L(V(Q_{11})) \times E(VVL) + S_L(V(Q_{12})) \times E(VL) \\ & + S_L(V(Q_{13})) \times E(L) + S_L(V(Q_{14})) \times E(M) \\ & + S_L(V(Q_{15})) \times E(H) + S_L(V(Q_{16})) \times E(VH) \\ & + S_L(V(Q_{17})) \times E(VVH)]/S_L(V(O_{11})) \\ & + S_L(V(O_{12})) + S_L(V(O_{13})) + S_L(V(Q_{14})) \\ & + S_L(V(Q_{15})) + S_L(V(Q_{15})) + S_L(V(Q_{17}))] \\ = {} & (0.85 * 30\% + 0.7 * 50\% + 0.6 * 70\%) \\ & /(0.85 + 0.7 + 0.6) = 47.67\%. \end{aligned}$$

Similarly, the *Expected Relative Market Shares* for the product Q_2 and Q_3 can be calculated as: $ERMS(Q_2) = 75.24\%$, and $ERMS(Q_3) = 59.17\%$, respectively.

According to the forecasted results: the *Expected Relative Market Shares* of products Q_1, Q_2 and Q_3 are 47.67, 75.24 and 59.17%, respectively; the *Expected Market Growth Rates* for products Q_1, Q_2 and Q_3 are 28.85, 7.6 and -23.48%, respectively. That is to say, comparing with current year, the forecasted year-over-year *Expected Market Growth Rate* of product Q_1 is *Good* and growing (G+). Likewise, comparing with current year, the forecasted year-over-year *Expected Market Growth Rates* of

Table 10 The market growth rate, relative market share and formulated strategy

Product no.	Expected market growth rate	Expected relative market share	Strategic position	Recommended strategy
Q_1	28.85%(G+)	47.67% (H+)	*Star*	Build or hold
Q_2	7.6% (G+)	75.24% (H+)	*STAR*	Build or hold
Q_3	−23.48%(B−)	59.17% (H−)	*CASH COW*	Harvest

product Q_2 is *Good* and increasing (G+). Yet, the forecasted year-over-year *Expected* Market Growth Rate of product Q_3 is *Bad* and decreasing (B−). The *Expected Market Growth Rate* and *Expected Relative Market Share* of each product will lead the product to a classification of one of the four Strategy Quadrants. Thus, the rules in Table 3 can be used to describe the classification rules of the domain knowledge involved in the product portfolio analysis. As shown in Table 10, the *Expected Market Growth Rates* and *Expected Relative Market Shares* of products Q_1, Q_2 and Q_3 lead to classification Rules of R1, R1 and R8, respectively. Take product Q_1 for example. Its "*Market Growth Rate*" is "High+" and "Market Share" is "Low−", then its identified "Strategy Quadrant" is "P*ROBLEM CHILD*" and the "Recommended Strategy" is "*Hold* or *Divest*".

In this paper, a new BCG product portfolio analysis method was proposed to the portfolio analysis panel using a forecasting vague grade sheet. The forecasting grades allocated to the products in the forecasting sheet can be regarded as a vague set, where each element in the universe of discourse belonging to a vague set is represented by a vague value in [0, 1]. As a best practice, a product portfolio analysis panel composed of expert and/or senior management is called on to forecast the company product's expected market growth rate and further to forecast he company product's expected relative market share based on their precious knowledge or experience of the firm or industry. The new rule based BCG growth-share matrix was then used for products' strategic market positions and for resources allocation. The case study results showed that the proposed new BCG Product portfolio analysis method has achieved the objectives of the research paper: (1) provided a collective evaluation method to the portfolio analysis panel to forecast company's product market size growth rate and relative market share growth rate, (2) recommended resource allocation strategy, depending on which strategic position the product fell. As a result, the new method has effectively and efficiently contributed in: (1) During measurement stage, forecasted (expected) market growth rate and relative market share; (2) During classification stage, classified the product into four market strategic positions: *Star*, *Cash Cow*, *Question Mark*, *or Dog*; (3) During recommendation stage, recommended resource allocation strategies: *Build*, *Hold*, *Harvest*, or *Divest*.

6 Conclusion

The main purpose of this study was to propose a new BCG method, which is a dynamic trend analysis method for identifying company products' strategic market positions and for formulating resources allocation strategy. The numerical case study results showed that the proposed BCG product portfolio analysis method has achieved the objectives of this study. It provides a collective forecasting method to the portfolio analysis panel for soliciting company products' expected market size growth rates and expected relative market share growth rates under vague and uncertain environment. It also provides recommendation strategies, depending on which strategy quadrants the products fall, to help company make resource allocation decisions related to their product portfolio strategies. Although we have proposed a new BCG analysis framework that will sustain future efforts, we believe that much work remains to be elaborated to the proposed BCG method. In the future, more score functions will be studied and used to the proposed BCG method for effective and efficient product portfolio analysis.

References

1. Wilson, R.M.S., Gilligan, C.: Strategic MARKETING MANAGEMENT: PLANNING, IMPLEMENTATION AND CONTROL, 3rd edn. Elsevier, Butterworth Heinemann, Oxford, England (2008)
2. Henderson, B.: Corporate Strategy. Abt Books Publisher, Cambridge (1979)
3. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**, 87–96 (1986)
4. Atanassov, K.T.: On Intuitionistic Fuzzy Sets Theory. Springer, Berlin, Heidelberg (2012)
5. Zadeh, L.A.: Fuzzy sets. Inf. Control. **8**, 338–356 (1965)
6. Gau, W.L., Buehrer, D.J.: Vague sets. IEEE Trans. Syst. Man, Cybernetics. **23**, 610–614 (1993)
7. Bustince, H., Burillo, P.: Vague sets are intuitionistic fuzzy sets. Fuzzy Sets Syst. **79**, 403–405 (1996)
8. Chen, S.M., Tan, J.M.: Handling multicriteria fuzzy decision-making problems based on vague set theory. Fuzzy Sets Syst. **67**, 163–172 (1994)
9. Lin, K.S., Chiu, C.C.: Multi-criteria group decision-making method using new score function based on vague set theory. In: Proceedings of International Conference on Fuzzy Theory and Its Applications (iFUZZY 2017, Pingtung, Taiwan), IEEE Xplore Digital Library. https://doi.org/10.1109/ifuzzy.2017.8311795
10. Boston Consulting Group: Perspectives on Experience, Boston. The Boston Consulting Group, MA (1974)
11. Chen, S.M.: Measures of similarity between vague sets. Fuzzy Sets Syst. **74**(2), 217–223 (1995)
12. Chen, S.M.: Similarity measures between vague sets and between elements. IEEE Trans. Syst., Man, Cybern.-Part B: Cybernetics. **27**(1), 153–158 (1997)
13. Henderson, B.: The Product Portfolio. https://www.bcg.com. Retrieved 8 Sept 2018

New Cost-Consequence FMEA Model for Information Risk Management of Safe And Secure SCADA Systems

Kuo-Sui Lin

Abstract Risk Priority Number (RPN) based Failure Mode and Effects Analysis (FMEA) can be used as a structured method to prioritize all possible vulnerable areas (failure modes) for review of safety and security in a supervisory control and data acquisition (SCADA) system. However, traditional RPN based FMEA has some inherent problems for risk management of information system. Therefore, the main purpose of this study was to propose a new cost-consequence FMEA model. It not only can recover traditional RPN-based FMEA problems, but also can evaluate, prioritize and correct safety and security of a SCADA system's failure modes. A numerical case study was conducted to demonstrate that the proposed cost-consequence FMEA model is not only capable of addressing FMEA's inherent problems but also is best suited for balancing monetary cost and risk consequence of failure modes in a SCADA system. It also facilitates to make better use of resources in optimizing cost and consequence of failure modes.

Keywords Equivalent annual cost · FMEA · Pattern recognition · SCADA · Vague set

1 Introduction

NIST SP 800-30 [1] defines risk as "a function of the likelihood of a given threat source exercising a particular potential vulnerability, and the resulting impact of that adverse event on the organization." Information risk management (IRM) is about identifying, assessing and prioritizing risks to keep information secure and available [2]. It is the process of determining which threats are of greater concern. At a high level risk analysis, this is accomplished by balancing exposure to risks against cost of mitigation and implementing appropriate control measures.

An industrial control system (ICS), or industrial automation and control system (IACS), is a generic term applied to hardware, firmware, communications, and

K.-S. Lin (✉)
Department of Information Management, Aletheia University, Taipei, Taiwan, R.O.C.
e-mail: au4234@mail.au.edu.tw

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_3

software used to perform vital monitoring and controlling functions of sensitive processes and enable automation of physical systems. The industrial control systems, which include supervisory control and data acquisition (SCADA) systems, distributed control systems (DCSs), and other smaller control system configurations such as skid-mounted Programmable Logic Controllers (PLCs), are often found in the industrial control sectors [3]. The largest subgroup of ICS is SCADA. In practice, SCADA is sometimes used interchangeably with ICS. ICSs or SCADAs operate critical infrastructure industries, such as electric power, oil and gas, water and wastewater, manufacturing, transportation, agriculture and chemical factories. Safety and security are key issues in SCADA systems. Design review and improvement of a safe and secure SCADA system aims at providing a structured technique to investigate potential failure modes and their effects as well as to improve overall safety and security risk of an SCADA system. In this study, safety is defined as the degree to which "accidental harm" is prevented, reduced and properly reacted to" and security is defined as the degree to which "malicious harm" is prevented, reduced and properly reacted to". Both definitions are taken from Firesmith [4]. Thus, the avoidance of harm from hazards and threats is dependent on both safety and security. In the context of a SCADA system, security risk is "a function of the likelihood of a given threat source exploiting a potential vulnerability and the resulting impact of a successful exploitation of the vulnerability" [3]. The threats can originate from the Internet, extranet, or other external networks, as well as from the operations and maintenance (O&M), and unauthorized access, etc. with the potential to result in workplace health and safety (WHS) incidents, environmental consequences and/or business disruption.

FMEA is a proactive risk management technique commonly used to identify and eradicate potential failures, problems and errors from a system, design, process or service [5]. Applications of FMEA for information security risk management have been widely studied in literature [6, 7]. In traditional RPN based FMEA, the overall risk of a failure mode is determined by calculating the *RPN* which is the multiplication of the three risk factors: Severity, Occurrence, Detection (i.e., $RPN = S * O * D$). Severity (impact or consequence) is the magnitude of harm that could be caused by a threat's exercise of a particular potential vulnerability. Detection or Detectability is the possible monitor or control measures that could mitigate and eliminate identified risk related to the vulnerable asset. Thus, FMEA is suited to analyze safety and safety scenario of SCADA systems. A reduction of any of the three risk factors will result in a reduction of safety and security scenario of SCADA systems.

However, traditional RPN based FMEA method has some inherent limitations as criticized by many researchers [8–11]. Some of the important disadvantages are restated as follows: (1) The RPN method assumes that the ratings of failure modes are crisp numerical values, but in many real world circumstances much information in the FMEA is vague and difficult to be precisely evaluated; (2) Different combinations of O, S and D may yield exactly the same RPN value, but their hidden risk implications may be totally different. For example, two different events with the (O, S, D) values of (2, 3, 2) and (4, 1, 3) respectively, have the same *RPN* value of 12. Yet, the hidden risks of these combinations are not necessarily identical. This may lead to a waste

of resources and time. Additionally, in some cases a high risk event may remain unnoticed; (3) The risk factors *O*, *S*, and *D* are assessed by discrete ordinal scales of measure, but the operation of multiplication is pointless on ordinal scales according to the measurement theory; (4) In the conventional FMEA, three risk factors *O*, *S*, *D* are considered with equal importance. Nonetheless, in the practical risk analysis the relative weightings of FMEA risk factors may be unequal. Besides, traditional evaluation method employs risk priority as the only criterion for correct actions of failure modes. A risk priority is high for correct action of certain failure mode does not mean that cost should be spent on that failure mode. That is to say, both risk priority and cost priority on correct action of failure mode must be considered simultaneously. However, there is a lack of previously stated research for existing information security risk management metrics on failure modes of a SCADA system.

Therefore, the purpose of this study was to propose a new cost-consequence FMEA model. It not only intends to recover traditional RPN based FMEA problems, but also to evaluate, prioritize and correct security of SCADA systems' failure modes, which are with imperfect and/or imprecise information.

In addition to solve the drawbacks of traditional RPN based FMEA problems, the novel contributions of this paper are: (1) to propose a vague set based RPN analysis for data collection that solicits rating values under vague and uncertain environment; (2) to propose a pattern recognition based FMEA method that classifies risk action priorities of failure modes, which will prompt accountable corrective action plans (CAPs) to improve design or process robustness of the evaluated SCADA system; (3) to propose a life cycle costing analysis approach that derives cost priorities of failure modes, which are with unequal budget life spans and are normally uncomparable; (4) to propose an action priority matrix that considers both risk priorities and cost priorities simultaneously and identifies action priority strategies for a SCADA system's failure modes.

2 Theoretical Background

2.1 Equivalent Annual Cost Approach for Life Cycle Costing Analysis

Equivalent annual cost (EAC) is used as a decision making tool in capital budgeting. When compare two or more investments on assets or CAPs with unequal life spans, EAC approach is suitable to convert the total cost of ownership (TCO) of an investment to an equivalent annual amount.

2.1.1 Life Cycle Cost

ISO 15686-5 [12] defined life cycle costing as an economic assessment considering all agreed projected significant and relevant cost flows over a period of analysis expressed in monetary value [12]. Life cycle cost (LCC) is the total cost of ownership (TCO) of an asset during the lifetime of its use and ownership [13]. The total monetary cost involves planning, purchase, operation and maintenance, and disposal or liquidation. It also includes those costs needed to achieve defined levels of performance, such as reliability, safety and availability. Within this study, it has been assumed that the total LCC is defined as [14]: $LCC_i = C_{INV} + C_{CM} + C_{PM} + C_{PL} + C_{REM}$, where C_{INV} is the cost of the investment; C_{CM} is the cost for corrective maintenance; C_{PM} is the cost for preventive maintenance; C_{PL} is the cost for production loss; C_{REM} is the disposal or remainder value for the investment i.

2.1.2 Equivalent Annual Cost

In capital budget decision making, equivalent annual cost (EAC) is the annual cost of owning, operating, and maintaining an investment on asset or CAP over its entire life. When compare two or more investments with unequal but repeating life spans, EAC is often used to convert the total life cycle cost of an investment on asset or i-th corrective action plan (CAP) to an equivalent annual amount [15] as: $NPV_i = \sum_{t=0}^{T} NCF_{it} \times (1+r)^{-t} = \sum_{t=0}^{T} EAC_i \times (1+r)^{-t}$, where Net Present Value (NPV_i) is the current value of all NCF_{it}; NCF_{it} stands for net cash flow(net future cash inflows and outflows) for the investment on i-th CAP for certain period of time t; T is the period during which the investments are used; r is the discount rate.

In this study, when compare two or more CAPs with unequal life spans, EAC approach is used to convert each TCO of CAP to an equivalent annual amount. As a first step, the NPV_i of the correct action investment (life cycle cost) to each failure mode FM_i need to be calculated as: $NPV_i = \sum_{t=0}^{T} \left[C_{it,INV} + C_{it,O} + C_{it,CM} + C_{it,PM} + C_{it,D}\right] \times (1+r)^{-t}$, where $C_{it,INV}$ is the cost of investment; $C_{it,O}$ is the net cash outflow for operation; $C_{it,CM}$ is the cash outflow for corrective maintenance; $C_{it,PM}$ is the cash outflow for preventive maintenance; $C_{it,D}$ is the disposal or remainder value for i-th failure mode; t is the number of periods from the present to the date of the out flow, and r is the discount rate; T is the period during which the investments are used. As a second step, the equivalent annual cost of an investment on i-th CAP is calculated by dividing the NPV_i by the "present value annuity factor $A_{t,r}$". Thus, EAC_i can be calculated using the following formula:

$$EAC_i = NPV_i / A_{t,r} = (NPV_i \times r) / \left[1 - (1+r)^{-T}\right], \tag{1}$$

where $A_{t,r} = [1 - (1+r)^{-T}]/r$; $A_{t,r}$ is the present value annuity factor $PVIFA\ (r, n)$ for interest rate r and n periods.

2.2 Vague Set Theory

Gau and Buehrer [16] proposed the concept of vague set (VS), where the grade of membership is bounded to a subinterval $[t_A(x), 1 - f_A(x)]$ of [0, 1]. The vague set (intuitionistic fuzzy set) has greater ability in depicting and handling uncertainty and imprecision due to insufficient knowledge. Relevant definition and operations of vague sets introduced in [16, 17] are briefly reviewed as follows.

Definition: Vague Sets
A vague set A in the universe of discourse X is characterized by a truth membership function, $t_A: X \to [0, 1]$, and a false membership function, $f_A: X \to [0, 1]$, where $t_A(x)$ is a lower bound of the grade of membership of x derived from the "evidence for x", and $f_A(x)$is a lower bound on the negation of x derived from the "evidence against x", and $0 \leq t_A(x) + f_A(x) \leq 1$. The grade of membership of x in the vague set is bounded to a subinterval $[t_A(x), 1 - f_A(x)]$ of [0, 1]. A vague value is simply defined as a unique element of a vague set. The interval $[t_A(x), 1 - f_A(x)]$ is called the grade of membership (vague value) of x in A. Let $f^*_A(x) = 1 - f_A(x)$, then the vague value $V_A(x) = [t_A(x), 1 - f_A(x)] = [t_A(x), f^*_A(x)]$.

The author [18, 19] introduced a transforming score function to transform vague values into comparable fuzzy membership values by presenting the hesitation margin as a distribution of triangular fuzzy number. The numerical membership values of the solicited vague values can be derived by the following transforming score function, shown as follows:

$$S_L(V_A(x)) = t_A(x)/2 + (1 - f_A(x))/2 = (t_A(x) + t_A * (x))/2. \quad (2)$$

2.3 Similarity Measure Between Vague Sets

In real world problems, a distance measure (similarity measure) can be used to represent the difference (similarity) between two objects expressed by corresponding fuzzy sets. Normalized Hamming distance is one of the most popular distance measures for fuzzy sets [20–22]. It can be used to derive similarity measure for vague sets.

Definition: Normalized Hamming distance between fuzzy sets
Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two fuzzy sets of the universe of discourse on $X = \{x_1, x_2, \ldots, x_n\}$. $\boldsymbol{A} = \mu_A(x_1)/x_1 + \mu_A(x_2) + \cdots + \mu_A(x_j)/x_j + \cdots + \mu_A(x_n)/x_n$, $\boldsymbol{B} = \mu_B(x_1)/x_1 + \mu_B(x_2) + \cdots + \mu_B(x_j)/x_j + \cdots + \mu_B(x_n)/x_n$, where $\mu_A(x_j)$ represents the membership value of x_j in A, $\mu_B(x_j)$ represents the membership value of x_j in $\boldsymbol{B}$. The normalized Hamming distance between vector $\boldsymbol{A}$ and vector $\boldsymbol{B}$ is just the sum of the two vector and is defined as: $dis_{nH}(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{n}\sum_{j=1}^{n} \left|\mu_A(x_j) - \mu_B(x_j)\right|$, where $\mu_A(x_j)$ represents the membership value of x_j in $\boldsymbol{A}$, $\mu_B(x_j)$ represents the membership value of x_j in $\boldsymbol{B}$.

2.3.1 Similarity Measure Between Vague Sets

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two vague sets of the universe of discourse on $X = \{x_1, x_2,\ldots, x_n\}$. $\boldsymbol{A} = [t_A(x_1), 1 - f_A(x_1)]/x_1 + [t_A(x_2),1 - f_A(x_2)]/x_2 + \cdots + [t_A(x_j),1 - f_A(x_j)]/x_j + \cdots + [t_A(x_n),1 - f_A(x_n)]/x_n$ and $\boldsymbol{B} = [t_B(x_1),1 - f_B(x_1)]/x_1 + [t_B(x_2),1 - f_B(x_2)]/x_2 + \cdots + [t_B(x_j), 1 - f_B(x_j)]/x_j + \cdots + [t_B(x_n), 1 - f_B(x_n)]/x_n$. The distance measure between the vague values $V_A(x_j) = [t_A(x_j), 1 - f_A(x_j)]$ and $V_B(x_j) = [t_B(x_j), 1 - f_B(x_j)]$ can be defined as the difference between the net membership degrees $M_A(x_j)$ and $M_B(x_j)$: $dis_L(V_A(x_j), V_B(x_j)) = M_A(x_j) - M_B(x_j)$, where $M_A(x_j)$ represents the net membership value of x_j in $\boldsymbol{A}$, $M_B(x_j)$ represents the net membership value of x_j in $\boldsymbol{B}$.

Based on Lin's membership score function [18, 19], the distance measure between the vague values $V_A(x_j)$ and $V_B(x_j)$ can be calculated as follows: $dis_L(V_A(x_j), V_B(x_j)) = M_A(x_j) - M_B(x_j) = |S_L(V_{Ai}(x_j)) - (S_L(V_{Bp}(x_j))| = |(t_A(x_j) + (1 - f_A(x_j))/2 - (t_B(x_j) + (1 - f_B(x_j))/2| = |(t_A(x_j) + (1 - f_A(x_j)) - (t_B(x_j) + (1 - f_B(x_j))|/2$, where $S_L(V_{Ai}(x_j))$ is the score of the vague value $V_{Ai}(x_j)$. Thus, the Hamming distance between vague sets $\boldsymbol{A}$ and $\boldsymbol{B}$ can be defined as follows:

$$\begin{aligned} dis_L(\boldsymbol{A}, \boldsymbol{B}) &= \frac{1}{n}\sum\nolimits_{j=1}^{n} \left|M_A(x_j)) - (M_{B1}(x_j))\right| \\ &= \frac{1}{n}\sum\nolimits_{j=1}^{n} \left|S_L(V_{Ai}(x_j)) - (S_L(V_{Bp}(x_j))\right| \\ &= \frac{1}{n}\sum\nolimits_{j=1}^{n} |(t_A(x) + (1 - f_A(x))/2 - (t_B(x) + (1 - f_B(x))/2| \\ &= \frac{1}{n} = \sum\nolimits_{j=1}^{n} |(t_A(x) - t_B(x))/2 - (f_A(x) - f_B(x))/2|. \end{aligned} \tag{3}$$

A similarity measure is assumed to be a dual measure to a distance measure expressed as $sim = 1 - dis$. Accordingly, the similarity between vague sets $\boldsymbol{A}$ and $\boldsymbol{B}$ can be defined as follows:

$$\begin{aligned} sim_L(\boldsymbol{A}, \boldsymbol{B}) &= 1 - \frac{1}{n}\sum\nolimits_{j=1}^{n} \left|M_A(x_j)) - (M_{B1}(x_j))\right| \\ &= 1 - \frac{1}{n}\sum\nolimits_{j=1}^{n} \left|S_L(V_{Ai}(x_j)) - (S_L(V_{Bp}(x_j))\right| \\ &= 1 - \frac{1}{n}\sum\nolimits_{j=1}^{n} |(t_A(x) + (1 - f_A(x))/2 - (t_B(x) + (1 - f_B(x))/2| \\ &= 1 - \frac{1}{n}\sum\nolimits_{j=1}^{n} |(t_A(x) - t_B(x))/2 - (f_A(x) - f_B(x))|/2. \end{aligned} \tag{4}$$

2.4 Failure Modes and Effects Analysis

FMEA is a team oriented analytic method to identify the functions of a product or a process and the associated potential failure modes, effects, and causes. The aim is to evaluate potential failure modes in order to assess the risk associated with the identified failure modes, and to prioritize the failure modes for identifying and carrying out corrective actions to address the most serious failure modes. Typical process of the FMEA is as follows: (1) Identify potential failure modes and their causes, (2) Evaluate the effects on the system of each failure mode, (3) Assess the risk associated with the identified failure modes, (3) Prioritize the failure modes, (4) Recommend corrective actions to reduce the risks associated with prioritized failure modes.

In 2017 FMEA Workshop, VDA (Verband der Automobilindustrie) and AIAG (Automotive Industry Action Group) agreed to harmonize and standardize a common set of FMEA requirements/expectations. A draft of the AIAG-VDA FMEA Handbook was released in November 2017 and a final version was issued in 2018 [23], that enable suppliers to have a single co-copyrighted AIAG-VDA FMEA Manual. The new AIAG-VDA FMEA handbook replaces "Fill-in-the-blank" with "Step Analysis" according to the new format. The new version adopts a structured approach and six-step implementation process: Step 1: Scope Definition, Step 2: Structure Analysis, Step 3: Function Analysis, Step 4: Failure Analysis, Step 5: Risk Analysis, Step 6: Optimization. The handbook offers logic on how to prioritize action priorities of failure modes. It recognizes that a higher RPN number may not be point to the correct item for the team to work on next. For example, there is no logic showing how a RPN rating of 90 should be prioritized over a RPN rating of 112. In the handbook, the severity rating is a measure associated with the most serious failure effect for a given failure mode of the function being evaluated. The severity rating shall be used to identify priorities of action plans relative to the scope of an individual FMEA and is determined without regard for occurrence or detection. As Table 1 shows, the manual reviews high severity ranks items first, and uses occurrence rating and detection rating to classify action priorities of failure modes.

3 Proposed New Cost-Consequence FMEA Model for Action Priorities of Failure Modes

Cost-consequence analysis is a form of health economic evaluation in which disaggregated costs and a range of different outcomes are presented to allow readers to form their own opinion concerning the relative importance of costs and outcomes. No specific preference for one costing approach or one outcome measure (as is the case for cost effectiveness analysis or cost-utility analysis) is made. Cost-consequences analysis reports a profile of health and non-health impacts for each intervention pro-

Table 1 Design FMEA action priority (AP)

AP	Justification for action priority—DFMEA
High	High priority due to safety and/or regulatory effects that have a High or Very High occurrence rating
High	High priority due to safety and/or regulatory effects that have a Moderate occurrence rating and High detection rating
High	High priority due to the loss or degradation of an essential or convenience vehicle function that has a Moderate occurrence rating and Moderate detection rating
Medium	Medium priority due to the loss or degradation of an essential or convenience vehicle function that has a Moderate occurrence and Low detection rating
Medium	Medium priority due to perceived quality (appearance, sound, haptics) with a Moderate occurrence and Moderate detection rating
Low	Low priority due to perceived quality (appearance, sound, haptics) with a Moderate occurrence and Low detection rating
Low	Low priority due to no discernible effect

gram. The reader or the decision maker has to form their own opinion concerning the relative importance of costs and outcomes [24].

Before proceed to the proposed FMEA model, the initial work to identify all possible failure modes FM_i, potential cause and frequency occurrence for each failure mode of the evaluated SCADA system was completed. Physical fallure modes of a SCADA system could be compromised by physical failure events. Similarly, cyber failure modes could be exploited by cyber threat events. If physical failure events or cyber treat events can be prevented by reducing the effects of the risk factors, the associated chain effects of the failure modes can be avoided. Physical failure risk of the failure modes of a SCADA system is a function of three factors (variables): the likelihood of failure occurrence (O), the effectiveness of failure Countermeasure (D), and the consequences of failure occurrence (S). The risk function assign three factors (variables) a rea number $R = f(S, O, D)$. Similarly, cyber failure risk of the failure modes of a SCADA system is a function of three factors (variables): the attack probability of failure occurrence (P), the effectiveness of failure Countermeasure (M), and the consequences of failure occurrence (C). The cyber security risk can be expressed by the following equation: $R = f(P, M, C)$, where: $R =$ Risk rating; $P =$ Attack probability rating which is a combination of Threat rating and Vulnerability rating of the vulnerability; $M =$ Countermeasure rating; the controls to the vulnerability; $C =$ Consequence rating; the impact consequence of the vulnerability. Once the threats, vulnerabilities, countermeasures and asset impact consequence can be categorized and characterized, the cyber security risk can be derived. Hence in this study, the three FMEA factors were adopted for safety and security risk evaluation of failure modes of a SCADA system.

3.1 The First Stage: Vague Set Based RPN Analysis for Risk Factor Ratings of Failure Modes

This stage is the data collection process for risk factor ratings of severity, occurrence and detection. In which FMEA team starts with understanding the system context and interviewing and reviewing various documents in order to evaluate the effectiveness of information risk management. During the RPN analysis stage, FMEA team solicited severity ratings, occurrence rating and detection ratings of failure modes. Given a set of evaluation team's rating values against the set of feature X for the data sample A_i, shown as: $\boldsymbol{A}_i = \{(x_1, r_{i1}), (x_2, r_{i2}),\ldots, (x_n, r_{in})\}$. Likewise, given a set of predefined rating values against the set of feature X for the data pattern $\boldsymbol{B}_i$ can be expressed as $\boldsymbol{B}_p$, shown as: $\boldsymbol{B}_p = \{(x_1, r_{p1}), (x_2, r_{p2}),\ldots, (x_, r_{pn})\}$. The rating value r_{ij} can be expressed by a linguistic rating or vague rating value $V_{Ai}(x_j)$. The rating value r_{pj} can be expressed by a linguistic rating or a vague rating value $V_{BP}(x_j)$. Thus, the rating vectors for the data sample and the rating vector for the data pattern can be expressed by the set $\boldsymbol{A_i}$ and $\boldsymbol{B}_p$, respectively.

Step 1.1: Soliciting linguistic Severity ratings $L(S_i)$ for FM_i

The *Severity* is related to the seriousness of the effects of a threat mode FM_i. In AIAG-VDA's FMEA handbook (AIAG-VDA [23], *Severity* is the dominant concern for recognizing action priorities of threat modes. In this study, four linguistic ratings are used for severity levels of threat modes, i.e., S_A (safety and/or regulatory effects), S_B (the loss or degradation of an essential function), S_C (loss or degradation of a convenience function), S_D (no discernible effect).

Step 1.2: Soliciting Occurrence vague rating and Detection vague rating of the j-th grade for the i-th failure mode

FMEA is a collective assessment process, and should be done by a team process manner. A FMEA team can employ the new polling method proposed by the author [18, 19] to solicit vague values $O(FM_{ij})$ and $D(FM_{ij})$ of the i-th failure mode FM_i ($i = 1,2,\ldots, m$) relative to the feature F_j ($j = 1,\ldots, n$). The individual degree assignments of each evaluator $k(k = 1,\ldots, q)$ voting in favor for the evaluated threat mode FM_i relative to the factor j is $t_{Ai}(x_{jk})$, voting against is $f_{Ai}(x_{jk})$, and in hesitation is $\pi_{Ai}(x_{jk})$, whereas: $t_{Ai}(x_{jk}) + f_{Ai}(x_{jk}) + \pi_{Ai}(x_{jk}) = 1$. Thus, the collective degree assignments of the evaluation team for the data samples $\boldsymbol{A}_i$ ($i = 1,\ldots,m$) relative to the factor $F_j (j = 1,\ldots, n)$ are obtained respectively as: $t_{Ai}(x_j) = \sum_{k=1}^{q} t_{Ai}(x_{jk})/q$, $f_{Ai}(x_j) = \sum_{k=1}^{q} f_{Ai}(x_{jk})/q$, $\pi_{Ai}(x_j) = \sum_{k=1}^{q} \pi_{Ai}(x_{jk})/q$. As a result, the rating value $r^{Ai}(x_j)$ for the evaluated data of failure mode FM_i on factor F_j can be solicited and expressed by a vague rating value: $V_{Ai}(x_j) = [t_{Ai}(x_j), 1 - f_{Ai}(x_j)] = [t_{Ai}(x_j), f^*_{Ai}(x_j)]$.

Step 1.3: Transforming numerical scores of the solicited vague Occurrence rating and vague Detection rating of the j-th grade for the i-th failure mode

In order to compute similarity measures between vague values of pattern data set $\boldsymbol{B}_p$ and sample data set $\boldsymbol{A}_i$, the solicited vague values $V_{Ai}(x_j)$ and $V_{BP}(x_j)$ must be

transformed into comparable numerical scores. Using Eq. 2, the numerical scores of the solicited vague values $V_{Ai}(x_j)$ and $V_{Bp}(x_j)$ can be derived by the following transforming score function, shown as follows:

$$S_L\big(V_{Ai}(x_j)\big) = t_{Ai}(x_j)/2 + \big(1 - f_{Ai}(x_j)\big)/2 = \big(t_{Ai}(x_j) + f^*_{Ai}(x_j)\big)/2.$$
$$S_L\big(V_{Bp}(x_j)\big) = t_{Bp}(x_j)/2 + \big(1 - f_{Bp}(x_j)\big)/2 = \big(\mathrm{tBp}(x_j) + f_{Bp} * (x_j)\big)/2.$$

3.2 The Second Stage: Pattern Recognition Based FMEA Method for Risk Action Priorities of Failure Modes

In this study, FMEA Action Priority created in AIAG-VDA FMEA handbook can be extended as a narrative rule set to recognize risk action priorities for failure modes. The business rules of risk action priority shown in Table 5 can be used to transform into more formal rules. The FMEA team can use the proposed pattern recognition method to recognize risk action priorities $RAP_i (i = 1, \ldots, m)$ for each of the identified failure modes FM_i $(i = 1,\ldots, m)$. Suppose there is a data sample $\boldsymbol{A}_i$ to be recognized against an attribute space $F = \{F_1, F_2,\ldots, F_n\}$, which is represented by the following formula: $\boldsymbol{A}_i = \{(x_j, r_{ij}) | x_j \in X\}$, $j = 1,2,\ldots, n$. Suppose that there exist p known patterns characterized against the attribute space $F = \{F_1, F_2,\ldots, F_n\}$, which is represented by the following formula: $\boldsymbol{B}_p = \{(x_j, r_{pj}) | x_j \in X\}$, $p = 1, 2,\ldots, t$. The following steps are proposed to solve fuzzy pattern recognition problems:

Step 2.1: Calculating the similarity measure sim$(\boldsymbol{A}_i, \boldsymbol{B_p})$

Two kinds of distance measures used in this study are attribute distance measure $dis(r_j^{Ai}, r_j^{Bp})$ and global distance measure $dis(\boldsymbol{A_i}, \boldsymbol{B_p})$. Feature distance measure is dissimilarity between feature vague values $V_{\boldsymbol{A}i}(x_j)$ and $V_{\boldsymbol{B}p}(x_j)$; Global distance measure is dissimilarity between data set $\boldsymbol{A_i}$ and $\boldsymbol{B_p}$.

Substep 2.1.1: Computing j-th attribute distance measure

The attribute distance is the distance measure on attribute between the two attribute ratings. Let $dis_L(r_j^{Ai}, r_j^{Bp})$ denote the j-th attribute distance measure between the data sample vector $\boldsymbol{A_i}$ and the pattern vector $\boldsymbol{B}_p$. r_j^{Ai} and r_j^{Bp} are the rating against attribute $\boldsymbol{F}_j$ in the data sample vector $\boldsymbol{A}_i$ and the pattern vector $\boldsymbol{B}_p$, respectively. In this study, the two ratings r_j^{Ai} and r_j^{Bp} can be expressed by the two vague values $V_{Ai}(x_j)$ and $V_{Bp}(x_j)$. Thus, the degree of the *j-th* attribute distance $dis_L(r_j^{Ai}, r_j^{Bp})$ between the data sample vector $\boldsymbol{A_i}$ and the pattern vector $\boldsymbol{B}_p$ can be evaluated as: $dis_L(r_j^{Ai}, r_j^{Bp}) = dis_L(V_{Ai}(x_j), V_{Bp}(x_j)) = |M_A(x_j)) - (M_{B1}(x_j))| = |S_L(V_{Ai}(x_j)) - (S_L(V_{B\mathrm{p}}(x_j))| = |(t_{Ai}(x_j) + (1 - f_A(x_j))/2 - (t_{Bp}(x_j) + (1 - f_{Bp}(x_j))/2|$.

Substep 2.1.2: Computing global similarity measure

The degree of similarity between the vague sets $\boldsymbol{A_i}$ and $\boldsymbol{B}_p$ can be evaluated by the global similarity function as follows: $sim_L(\boldsymbol{A}_i, \boldsymbol{B}_p) = 1-(\sum_{j=1}^{n} w_j \times dis_L(r_j^{Ai}, r_j^{Bp}))/\sum_{j=1}^{n} w_j$, where w_j is the attribute weights allocated to each attribute reflecting importance of the corresponding attribute. Suppose the weightings are equally weighted, and each of the global similarity measures can be calculated as

$$\begin{aligned} sim_L(\boldsymbol{A}_i, \boldsymbol{B}_p) &= 1-\frac{1}{n}\sum\nolimits_{j=1}^{n} dis_L(V_{Ai}(x_j), V_{Bp}(x_j)) \\ &= 1-\frac{1}{n}\sum\nolimits_{j=1}^{n} \left|M_A(x_j)) - (M_{B1}(x_j))\right| \\ &= 1-\frac{1}{n}\sum\nolimits_{j=1}^{n} \left|S_L(V_{Ai}(x_j)) - (S_L(V_{Bp}(x_j))\right| \\ &= 1-\frac{1}{n}\sum\nolimits_{j=1}^{n} \left|(t_{Ai}(x_j) + (1 - f_A(x_j))/2 - (t_{Bp}(x_j) + (1 - f_{Bp}(x_j))/2\right| \\ &= 1-\frac{1}{n}\sum\nolimits_{j=1}^{n} \left|(t_A(x_j) + t_A * (x_j))/2 - (t_B(x_j) + t_B * (x_j))/2\right|. \end{aligned} \tag{5}$$

Step 2.1.3: Compare global similarity measures and select the largest one

According to the recognition principle of maximum degree of similarity, the process of assigning $\boldsymbol{B_p}$ to $\boldsymbol{A}_i$ is described by:

$$G_j = arg \max_{1\le p\le t} sim_L\left(\boldsymbol{A}_i, \boldsymbol{B}_p\right). \tag{6}$$

Compare $sim_L(\boldsymbol{A_i}, \boldsymbol{B}_p)$ *for* $p = 1,2,\ldots, t$ and select the largest one, denoted by $sim_L(\boldsymbol{A_i}, \boldsymbol{B}_p^*)$, *from* $sim_L(\boldsymbol{A_i}, \boldsymbol{B}_p)$ $(p = 1,2,\ldots, t)$. A pattern $\boldsymbol{B}_p^*$ can be derived such that $sim_L(\boldsymbol{A_i}, \boldsymbol{B}_p^*) = max\{sim_L(\boldsymbol{A_i}, \boldsymbol{B}_p^*) | \, p = 1,2,\ldots, t\}$. Then the data sample $\boldsymbol{A_i}$ belongs to the pattern $\boldsymbol{B}_p^*$.

Step 2.2: Select the next data sample to proceed until all data samples have been classified.

3.3 The Third Stage: Life Cycle Costing Analysis Approach *for* Cost Priorities of Failure Modes

Step 3.1: Exploit correct actions for failure modes

The recommended CAPs for risk reduction fall into three categories: design improvement plan, operations and maintenance (O&M) improvement plan, and emergency preparedness plan (EPP). In this step, recommendations of risk action priorities $RAP_i(i = 1,\ldots, m)$ to enhance the performance of failure modes are proposed, which

Table 2 Linguistic variables and their value range

Linguistic variable—cost priorities	Value ranges
Low cost priority	25% < annual budget
Medium cost priority	25% ≤ annual budget ≤ 75%
High cost priority	75% < annual budget

may include preventive actions and corrective actions. By implementing each of the actions associated with RAP_i, the *RPN* factors for each failure mode FM_i can be reduced.

Step 3.2: Calculate equivalent annual cost for correct actions of failure modes

Investment to each CAP of failure modes may have different useful lives. Given such a situation, the concept of equivalent annual cost is useful for handling the unequal life span.

Step 3.3: Transform annual correction budget into cost priorities

As shown in Table 2, the linguistic variables *Low Cost Priority*, *Medium Cost Priority* and *High Cost Priority* are defined as "Less Than 25%", "Between 26 and 75%", and "Greater Than 75%" of the annual total correction budget, respectively. In which, the annual total correction budget for each CAP of failure mode is calculated using formula of equivalent annual cost (EAC) (Eq. 1).

3.4 The Fourth Stage: Action Priority Matrix for SCADA's Failure Modes

Prioritizing CAPs for failure modes is a critical decision making point for information risk management of a SCADA system. Traditional evaluation method employs risk priority as the only criterion for CAPs of failure modes. Focusing on just the risk priority could lead to very different conclusion, or even opposite result. A risk priority is high for CP of certain failure mode does not mean that cost should be spent on that failure mode. Similarly, looking only at the cost priority and focusing only on fixing the failure mode at relatively low cost may be of little value, if the failure mode is not with high risk priority. That is to say, both risk priority and cost priority on CAP of failure mode must be considered simultaneously. However, there is a lack of previously stated research for existing information security risk management metrics on failure modes. Therefore, in this study, an action priority matrix was proposed that considers both risk priority and cost priority simultaneously and identifies action priority strategies for failure modes needing correct actions. The risk-cost action priority matrix intends to prioritize different correct action priority strategies for failure modes. This can result in better allocating organizational resources and deploying CAPs for a secure and safe SCADA system. As shown in Fig. 1, the risk priority

	High Risk Priority	Medium Risk Priority	Low Risk Priority
High Cost Priority	The 3rd Action Priority	The 6th Action Priority Strategy	The 9th Action Priority Strategy
Medium Cost Priority	The 2nd Action Priority Strategy	The 5th Action Priority Strategy	The 8th Action Priority Strategy
Low Cost Priority	The 1st Action Priority Strategy	The 4th Action Priority Strategy	The 7th Action Priority Strategy

Fig. 1 Risk-cost action priority matrix

and cost priority are reviewed simultaneously to classify action priority strategies for failure modes needing correct actions.

4 Numerical Case Study

4.1 Implementation of the Case Study

Following the procedure of the new FMEA model, a numerical case study was conducted to demonstrate the efficiency and effectiveness of the proposed method. In this case study, risk evaluation for the SCADA system was carried out by dividing the whole system into its sub-units. Each sub-unit was further divided up to component level and failure mode of each component was discussed in detail depending on safety or security concerns from the harm from hazard or threat. Before proceed to the case study, the initial work to identify all possible failure modes *FMi*, potential cause and frequency occurrence for each failure mode of the evaluated SCADA system was completed. A total of 125 failure modes were identified. The *Severity* ratings can be solicited and described by linguistic rating values, The *Occurrence* ratings and *Detection* ratings can be solicited and expressed by vague rating values. The *Occurrence* vague ratings and *Detection* vague ratings for the failure modes FM_i ($i = 1,\ldots, 125$) are then transformed into comparable scores and summarized in Table 3.

Table 3 Transformed rating scores for failure modes

No. of failure mode	Severity rating score $S(FM_i)$	Occurrence rating score $O(FM_i)$	Detection rating score $D(FM_i)$
FM_1	S_B	0.741	0.932
FM_2	S_A	0.779	0.851
FM_3	S_A	0.851	0.510
⋮	⋮	⋮	⋮
FM_{125}	S_B	0.757	0.834

Table 4 Narrative FMEA Rules for Risk Action Priority

RAP rules	Action priority	Justification for action priority
R_1	*High*	*High* priority due to "safety and/or regulatory effects" that has a *High* occurrence rating
R_2	*High*	*High* priority due to "safety and/or regulatory effects" that has a *High* detection rating
R_3	*High*	*High* priority due to "the loss or degradation of an essential function" that has a *High* occurrence rating
R_4	*High*	*High* priority due to "the loss or degradation of an essential function" that has a *High* detection rating
R_5	*Moderate*	*Moderate* priority due to "the loss or degradation of an essential function" that has a *Moderate* occurrence rating
R_6	*Moderate*	*Moderate* priority due to "the loss or degradation of an essential function" that has a *Moderate* detection rating
R_7	*Low*	*Low* priority due to "the loss or degradation of an essential function" that has a *Low* occurrence rating and *Low* detection rating
R_8	*Low*	*Low* priority due to "no discernible effect"

Table 5 Formal rules of risk action priority (*RAP*) for failure modes

Action priority rules	Severity C_1:S(FM_i)	Occurrence C_2:O(FM_i)		Detection C_3:D(FM_i)		RAP
R_1	S_A	H	0.72	–	–	*High*
R_2	S_A	–	–	H	0.72	*High*
R_3	S_B	H	0.72	–	–	*High*
R_4	S_B	–	–	H	0.72	*High*
R_5	S_B	M	0.52	–	–	*Moderate*
R_6	S_B	–	–	M	0.52	*Moderate*
R_7	S_B	L	0.32	L	0.32	*Low*
R_8	S_C	–	–	–	–	*Low*

Inspired by AIAG-VDA's FMEA Action Priority [23], the narrative rules of risk action priority shown in Table 4 are proposed in this study for pattern recognition.

The above narrative rules of risk action priorities are then transformed into more formal rules, as shown in Table 5.

The formal rules shown in Table 5 can be rewritten as:

R_1: IF{(C_1, S_A) And $(C_2, 0.72)$ And $(C_3, -)$}, Then $RAP = High$,

R_2: IF{(C_1, S_A) And $(C_2, -)$ And $(C_3, 0.72)$}, Then $RAP = High$,

(continued)

(continued)

R_3: IF{(C_1, S_B) And (C_2, 0.72) And (C_3, –)}, Then $RAP = High$ R_4: IF{(C_1, S_B) And (C_2, –) And (C_3, 0.72)}, Then $RAP = High$, R_5: IF{(C_1, S_B) And (C_2, 0.52) And (C_3, 0.52)}, Then $RAP = Moderate$,
R_6: IF{(C_1, S_B) And (C_2, –) And (C_3, 0.52)}, Then $RAP = Moderate$,
R_7: IF{(C_1, S_B) And (C_2, 0.32) And (C_3, 0.32)}, Then $RAP = Low$,
R_8: IF{(C_1, S_C) And (C_2, –) And (C_3, –}, Then $RAP = Low$.

A set of linguistic level grades are used to represent the rating grades regarding a failure mode portfolio FM, where $FM \in \{FM_1, FM_2, \ldots, FM_i, \ldots, FM_{125}\}$. Assume that the set of level grade assigned to evaluate the strength of the risk factors of the FM_i is G, where $G_j \in \{G_1, G_2, \ldots, G_k, \ldots, G_p\}$ and $0\% \le m_{1j} \le E(G_j) \le m_{2j} \le 100\%$. Then the expected rating values against the j-th linguistic level grade G_j are evaluated for each failure mode FM_i as follows:

$$E(G_j) = (1-\lambda) \times m_{1j} + \lambda \times m_{2j}, \tag{7}$$

where $\lambda \in [0, 1]$ denotes the optimism index determined by the evaluator, and $E(G_j)$ is the expected rating values against the j-th linguistic grade G_j. If $0 \le \lambda \le 0.5$, the evaluator is a pessimistic evaluator. If $\lambda = 0.5$, the evaluator is a normal evaluator. If $0.5 \le \lambda \le 1.0$, the evaluator is an optimistic evaluator. In this study, the three level grades used for the evaluation of the system's rating values are: $G = \{G_1, G_2, G_3\} = \{Low(L), Moderate(M), High(H)\}$. The ranges of level grades and the degrees of the level grades are shown in Table 6. The optimism index λ determined by the evaluator is 0.60 (i.e., $\lambda = 0.60$). Based on Eq. 7, the expected rating values $E(G_j)$ of the assigned linguistic level grade G_j for the i-th failure mode is calculated as follows: $E(G_3) = (1–0.60) \times 60\% + 0.60 \times 100\% = 84\%$, $E(G_2) = (1 - 0.60) \times 30\% + 0.60 \times 70\% = 54\%$, $E(G_1) = (1 - 0.60) \times 0\% + 0.60 \times 40\% = 24\%$.

In the first stage of the case study, FM_{125} was taken as an illustrative sample data for describing proposed pattern recognition method. The transformed rating scores $O(FM_{125})$ and $D(FM_{125})$ for sample data FM_{125} are calculated as 0.757 and 0.834, respectively, as shown in Table 3. Also in Table 3, the linguistic rating S_B and vague scores for FM_{125} are used to represent the data sample vector $\boldsymbol{A}_{125}$, shown as: $\boldsymbol{A}_{125} = S_B/x_1 + 0.757/x_2 + 0.834/x_3$. In which, "Loss or degradation of an essential function" is the dominant concern of the FM_{125} sample data. Consequently, rule R_3 through R_7 are screened to recognize the best suited pattern of the sample data $\boldsymbol{A}_{125}$. Thus, the data patent of the rules R_3 through R_7 are represented as: $\boldsymbol{B}_3 = S_B/x_1 + 0.84/x_2$,

Table 6 Level grade and corresponding degree of level grade

Level grade	Ranges of level grade (%)	Degrees of level grade (%)
Low (L)	0–40	24
Moderate (M)	30–70	54
High (H)	60–100	84

	High Risk Priority	Medium Risk Priority	Low Risk Priority
High Cost Priority	The 3rd AP FM_{125} ★	The 6th AP	The 9th AP
Medium Cost Priority	The 2nd AP	The 5th AP FM_{124} ★ FM_{15} ★	The 8th AP
Low Cost Priority	The 1st AP FM_{1} ★	The 4th AP	The 7th AP FM_{123} ★

Fig. 2 Results of cost-risk action priority matrix

$\boldsymbol{B}_4 = S_B/x_1 + 0.84/x_3$, $\boldsymbol{B}_5 = S_B/x_1 + 0.54/x_2$, $\boldsymbol{B}_6 = S_B/x_1 + 0.54/x_3$, $B_7 = S_B/x_1 + 0.24/x_2 + 0.24/x_3$, respectively.

In the second stage, by applying the proposed similarity function (Eq. 5), the matching similarity $sim_L(\boldsymbol{A_{125}}, \boldsymbol{B}_p)(p = 3, 4, 5, 6, 7)$ can be calculated as: $sim_L(\boldsymbol{A_{125}}, \boldsymbol{B}_3) = 1 - |0.757 - 0.84| = 0.917$, $sim_L(\boldsymbol{A_{125}}, \boldsymbol{B}_4) = 1 - |0.834 - 0.84| = 0.994$, $sim_L(\boldsymbol{A}_{125}, \boldsymbol{B}_5) = 1 - |0.757 - 0.54| = 0.783$, $sim_L(\boldsymbol{A_{125}}, \boldsymbol{B_6}) = 1 - |0.834 - 0.54| = 0.706$, $sim_L(\boldsymbol{A}_{125}, \boldsymbol{B}_7) = 1 - ((0.757 - 0.24) + (0.834 - 0.24))/2 = 0.445$. The larger values $sim_L(\boldsymbol{A_{125}}, \boldsymbol{B}_p)$ is, the higher the degree of similarity matching become. According to the recognition principle of maximum degree of similarity (Eq. 6), it can be observed that data sample A_{125} should be classified to pattern B_4, which states that "Rule 4: *High* priority due to the "loss or degradation of an essential function" that has a *High* detection rating."

In the third stage, when compare investments on CAPs of failure modes with unequal life spans, EAC approach is used to convert the total cost of ownership (TCO) of correction budget investments into respective equivalent annual amount (Eq. 1). The derived annual amounts for CAPs of failure modes are then transformed into their matching cost priorities.

In the fourth stage, the proposed action priority matrix was used that considers both risk priority and cost priority simultaneously and identifies action priorities for CAPs of failure modes. As shown in Fig. 2, the table classifies *High* cost priority and *High* risk priority simultaneously for CAP of failure mode FM_{125}. Consequently, the 3rd action priority for failure mode FM_{125} is identified and marked in the cost-risk action priority matrix. Similarly, the matching action priorities for failure modes FM_{124}, FM_{123}, FM_1, etc., are also identified and marked in the matrix.

4.2 Findings and Discussion on the Case Study

The proposed model is not only capable of addressing FMEA's inherent problems but also is best suited for balancing monetary cost and risk consequence of failure modes in a SCADA system. In the first stage, a vague set based RPN analysis was performed to solicit *Severity* ratings, *Occurrence* ratings and *Detection* ratings of failure modes.

In the second stage, a new pattern recognition based FMEA method was proposed to derive risk priorities of identified failure modes. In the third stage, a life cycle costing analysis approach was proposed to derive cost priorities of identified failure modes with unequal budget life spans. In the fourth stage, a cost-risk action priority matrix was proposed that considers above risk priorities and cost priorities simultaneously to identify action priorities for identified failure modes.

The result from this case study confirms the applicability and benefits of the proposed model: (1) Under vague and uncertain situations, FMEA team members are hesitant in expressing their diverse risk assessments over failure modes. To manage such situations, the novel risk priority approach based on vague set theory was contributed to enhance the assessment capability of FMEA under vague and uncertain environment. (2) The misconception of action priorities judged from different combinations of *O*, *S* and *D* are also prevented because of a new FMEA method was proposed to prioritize risk correct actions of identified failure modes of the SCADA system. (3) In view of the ordinal scale multiplication problem, the vague set theory is useful for soliciting numerical scores to *SOD* risk factors of the failure modes identified by the FMEA team. (4) The proposed FMEA model also can provide relative importance weightings of risk factors *O*, *S*, and *D* to prevent hidden risks of identical combinations. (5) The life cycle costing analysis approach was used to derive cost priorities of failure modes, which are with unequal budget life spans and are normally uncomparable. (6) The cost-risk action priority matrix was proposed that considers both risk priorities and cost priorities simultaneously and identifies action priorities for failure modes needing correct actions. (7) The results of the case study demonstrated that the proposed model is useful for recognizing risk priorities and for deploying cost priorities of failure modes.

Consequently, the proposed action priority matrix is an effect and efficient strategic analysis tool for better allocating organizational resources and deploying accountable CAPs for failure modes. Theoretically, it provides a semi-quantitative analysis of a SCADA system's failure modes in the early design phases. In practical implications, it enables managers and designers to estimate the value for monetary cost of a new CAP for a specific failure mode. It also facilitates to make better use of resources in optimizing cost and consequence of failure modes for information risk management of safe-secure SCADA systems.

5 Conclusions

For a safe-secure SCADA system, physical failure modes or cyber failure modes could be compromised by physical failure events or cyber threat agents (threat events). In this study, a new cost-consequence FMEA model is proposed, in order to recover inherent limitations of traditional RPN based FMEA and to evaluate, prioritize and correct safety and security in a SCADA system's failure modes. An action priority matrix was proposed that considers both risk action priorities and cost priorities simultaneously and identifies action priority strategies for failure modes needing

correct actions. Finally, a numerical case study was conducted to demonstrate that the proposed FMEA model is not only capable of addressing traditional FMEA's inherent problems but also is effective and efficient to be used as the basis for qualitative high level risk management and continuous improvement of a safe and secure SCADA system in the early design phases. Thus, the proposed FMEA model can contribute to better allocating organizational resources and deploying accountable CAPs for the development of safe-secure SCADA systems. In the future, preliminary scenarios of consequential loss coverage provided by existing cyber insurance policies can be reflected and augmented in the narrative rules for recognizing action priorities depending on the size and complexity of the evaluated SCADA system. Besides, more low level information and data relevant to failure modes will be studied and collected to conduct further case studies for information risk management of Safe and Secure SCADA systems.

References

1. NIST: Special Publication 800-30, Revision 1, Guide for Conducting Risk Assessments, September. National Institute of Standards and Technology, Gaithersburg, MD (2012)
2. Sutton, D.: Information Risk Management. BCS Learning & Development Limited, UK, Swindon (2015)
3. NIST: Special Publication 800-82 Revision 2, Guide to Industrial Control Systems Security. National Institute of Standards and Technology, Gaithersburg, MD (2015)
4. Firesmith, D.G.: Common Concepts Underlying Safety, Security, and Survivability, Technical note CMU/SEI-2003-TN-033, Software Engineering Institute, Pittsburgh. Carnegie Mellon University, PA (2003)
5. Stamatis, D.H.: Failure Mode and Effect Analysis: FMEA from Theory to Execution, 2nd edn. ASQ Quality Press, New York (2003)
6. Asllani, A., Lari, A., Lari, N.: Strengthening information technology security through the failure modes and effects analysis approach. Int. J. Qual. Innovation **4**(5), 1–14 (2018)
7. Silva, M.M., de Gusmão, A.P.H., Poleto, T., e Silva, L.C., Costa, A.P.C.S.: A multidimensional approach to information security risk management using FMEA and fuzzy theory. Int. J. Inf. Manage. **34**(6), 733–740 (2014)
8. Bowles, J.B., Pelaez, C.E.: Fuzzy logic prioritization of failures in a system failure modes, effects and criticality analysis. Reliab. Eng. Sys. Safety **50**(2), 203–213 (1995)
9. Chang, K.H., Cheng, C.H., Chang, Y.C.: Reprioritization of failures in a silane supply system using an intuitionistic fuzzy set ranking technique. Soft. Comput. **14**(3), 285–298 (2010)
10. Chin, K.S., Wang, Y.M., Poon, G.K.K., Yang, J.B.: Failure mode and effects analysis by data envelopment analysis. Decis. Support Syst. **48**(1), 246–256 (2009)
11. Sankar, N.R., Prabhu, B.S.: Modified approach for prioritization of failures in a system failure mode and effects analysis. Int. J. Qual. Reliab. Manag **18**(3), 324–335 (2001)
12. ISO 15686-5: Buildings and Constructed Assets-Service-Life Planning-Part 5: Life-cycle Costing, International Organization for Standardization standard (2017)
13. OGC: Whole Life Costing and Cost Management, Achieving Excellence in Construction, Procurement Guide, Number 07. Office of Government Commerce (2007)
14. Nilsson, J., Bertling, L.: Maintenance management of wind power systems using Condition monitoring systems-life cycle cost analysis for two case studies in the Nordic system. IEEE Trans. Energy Convers. **22**(1), 223–229 (2007)
15. Kogan, A.: The criticism of net present value and equivalent annual cost. J. Adv. Res. Law Econ. **1**(9), 15–22 (2014)

16. Gau, W.L., Buehrer, D.J.: Vague sets. IEEE Trans. Syst. Man Cybern. **23**, 610–614 (1993)
17. Chen, S.M., Tan, J.M.: Handling multicriteria fuzzy decision-making problems based on vague set theory. Fuzzy Sets Syst. **67**(2), 163–172 (1994)
18. Lin, K.S., Chiu, C.C.: Multi-criteria group decision-making method using new score function based on vague set theory. In: 2017 International Conference on Fuzzy Theory and Its Applications (iFUZZY 2017), pp. 1–6, Pingtung, Taiwan (2017)
19. Lin, K.S.: Efficient and rational multi-criteria group decision making method based on vague set theory. J. Comput. Accepted 11/18/2018. (in press)
20. Bagchi, S.: Performance and quality assessment of similarity measures in collaborative filtering using mahout. Procedia-Procedia Comput. Sci. **50**, 229–234 (2015)
21. Kaufmann, A., Gupta, M.M.: Introduction to Fuzzy Arithmetic Theory and Applications. Van Nostrand Reinhold, New York (1991)
22. Szmidt, E., Kacprzyk, J.: Distances between intuitionistic fuzzy sets. Fuzzy Set Syst. **114**, 505–518 (2000)
23. AIAG-VDA: Failure Mode and Effect Analysis (FMEA) Handbook, 1st edn. (2018)
24. Hillger, C.: Lifestyle and health determinants. In: Kirch, W. (ed.) Encyclopedia of Public Health. Springer, New York (2008)

A Dynamic Privacy Preserving Authentication Protocol in VANET Using Social Network

Syed Asad Shah, Chen Gongliang, Li Jianhua and Yasir Glani

Abstract Vehicular ad-hoc network (VANET) facilitate users with road safety and improve traffic efficiency by enabling vehicles to broadcast traffic information used by emergency electronic brake light and platooning. Communication in an open access environment makes authentication of the broadcasted message and privacy of the sender a challenging and critical issue. The symmetric cryptography based protocols have high throughput and are short in size but failed to ensure non-repudiation. However, asymmetric cryptography protocols based on the certificate revocation list (CRL) causes the storage overhead and are computationally expensive. In this paper, we propose a dynamic privacy-preserving authentication protocol using a hybrid combination of symmetric advanced encryption standard (AES) and asymmetric elliptic curve cryptography (ECC). We also propose a dynamic region topology algorithm to ensure the message gets delivered to desired vehicles in a region using the social network. The security and performance analysis shows that our proposed protocol is privacy-preserving, secure, light-weight and efficient.

1 Introduction

According to international road traffic and accident database (IRTAD) road safety annual report [1] 2018, the average number of road deaths in 32 IRTAD member countries have increased to 80648 per year. According to the report, 20–30% of fatal road crashes are due to over-speed driving. Distracted driving, increasing population,

S. A. Shah (✉) · C. Gongliang · L. Jianhua · Y. Glani
School of Information Security Engineering, Shanghai Jiao Tong university, Shanghai, China
e-mail: asad_shh786@sjtu.edu.cn

C. Gongliang
e-mail: chengl@sjtu.edu.cn

L. Jianhua
e-mail: lijh888@sjtu.edu.cn

Y. Glani
e-mail: yasirglani@sjtu.edu.cn

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_4

bad road condition, traffic rules violation and lack of communication are the main causes of road accidents. To overcome these fatal road crashes, Intelligent Transport System (ITS) [2] is introduced named Vehicular Ad-Hoc Network (VANET). VANET is evolved by applying principles of the mobile ad-hoc network (MANET) with characteristics like changing topology and high mobility [3]. VANET provide vehicles a safe and reliable environment by reducing road congestion [4]. A typical VANET communication is based on vehicle to vehicle (V2V) and vehicle to infrastructure(V2I). Vehicles communicate with each other using dedicated short-range communication (DSRC) technique [5, 6] with the help of the On-Board Unit (OBU) and Road-Side Unit (RSU).

Vehicle broadcast both critical and non-critical messages during communication. Non-critical messages include weather forecast, car parking, and entertainment etc. Critical messages may contain road safety information including road conditions warning, post-crash warning, lane changing assistance, intersection collision avoidance, and traffic assistance. Several applications used this information such as electronic brake light, road transportation emergency service and platooning. The message broadcast in an open environment raises privacy and security issues related to vehicle identification and location. Someone may use this information to threaten the owner of the vehicle. However, there is a risk involved that if the attacker eavesdrop or spoof bogus messages to create a false impression of the road congestion for his own good sake that may lead to an accident. VANET does not override all solution of security attack related to MANET due to its unique features like low volatility and high mobility. The fundamental security requirements in VANET include authentication (vehicle, broadcasted message) and privacy of the vehicle. Few VANET security attacks described in [7, 8] are discussed below.

- Modification Attack: In this attack, the content of the message are modified by the attacker that may lead to an accident.
- Location Tracking Attack: The attacker tracks the location of the vehicles may later use this information to harass the user.
- Impersonation Attack: The fake identity used by the attacker in order to pretend to be another vehicle.
- Replay Attack: The authentic information is maliciously repeated or delayed.
- False information Attack: The attacker broadcasts bogus or false information.
- Denial of Service Attack: Attackers reduces network performance by injecting false messages.

Many symmetric and asymmetric PKI based schemes have been proposed to address the security issues in VANET. But each scheme has its own limitation. Symmetric cryptography based protocols have high throughput and short in size but failed to ensure non-repudiation. In asymmetric cryptography based protocols vehicles are registered certificate authority (CA). All registered vehicles get the certificate from CA which include identity (ID) of CA and vehicles public key. The certificate is attached with the beacon signal for vehicle authentication. Receiving vehicle verifies the sender certificate by checking a very large file of certificate revocation list (CRL).

This paper propose a dynamic efficient lightweight privacy-preserving authentication protocol using a hybrid combination of symmetric AES and asymmetric ECC to overcome the aforesaid issues regarding inter-vehicle communication (IVC) using the social network. We also propose a lightweight algorithm to ensure the message gets delivered to desired vehicles in a region.

The rest of the paper is organized as follows: Sect. 2 discusses the related work and background. Section 3 presents our proposed protocol. Section 4 presents the security analysis and simulation results. Finally, we conclude in Sect. 5.

2 Related Work and Background

In VANET, authentication plays a vital role in vehicles communication. Most of the privacy-preserving authentication schemes are based on symmetric and asymmetric PKI. Raya et al. [9] proposed a PKI based scheme to provide privacy in which each vehicle needs to save a lot of certificates and public keys. The CRL check causes the denial of service (DOS) attack due to transmission and computation overheads. Varshney et al. [10] have proposed a protocol in which they used a digital certificate to provide VANET security. Tangade et al. [11] proposed a hybrid cryptographic scheme based on asymmetric PKI and symmetric HMAC code during V2I and V2V communication to provide scalable and privacy-preserving authentication. However, this scheme required the pervasive deployment of RSU. Sun et al. [12] have proposed an authentication scheme in which certificates used a single public key for a long period without considering location and privacy issues.

Rhim et al. [13] discussed efficient message authentication schemes using the message authentication code (MAC). But these schemes are vulnerable against a replay attack. Hussain et al. [14] proposed a hierarchical pseudonymous based approach in which they receive pseudonym from both CA and RSU. The major drawback of their scheme was the pervasive deployment of RSU. Lu et al. [15] propose a privacy-preserving scheme in which RSU distribute short-time pseudonyms keys to vehicles. This scheme also required a larger amount of RSU. The schemes focused in [16–18] used identity (ID) based cryptography where public keys are recognizable identity concealed with the help of pseudonym and therefore it causes pseudonym management overhead. Jiang et al. [19] proposed a scheme based on the hash message authentication code to prevent the time-consuming CRL checking procedure. Anirudh et al. [20] have proposed an efficient message authentication protocol (Mavanet) with QR encryption and decryption algorithm using the social network. The connection between sender and receiver is based on their social tie-ups.

Our proposed protocol neither requires the pervasive deployment of RSU nor the distribution and management of CRL checking. We used a hybrid combination of symmetric AES and asymmetric ECC to ensure efficient privacy-preserving authentication.

3 The Proposed Protocol

In our proposed protocol, vehicle first needs to register with the social network in order to receive and send the broadcasted message. First, we discuss the preliminaries and dynamic network architecture of our proposed protocol.

3.1 Preliminaries

We considered the geographical map of the city is divided into many smaller regions. The size of each region is based on traffic consumption helps to allocate resources as shown in Fig. 1. Each region has a unique identifiers region identity Reg-Id, center coordinates (Lat_C, Lon_C) and radius (R) stored in vehicle OBU. As a vehicle starts a journey OBU runs the Algorithm 1 and assigns vehicle the Reg-Id based on its location.

3.2 Network Topology

Our proposed protocol network architecture is a hybrid combination of Piconet and Scatternet. In which, one vehicle is elected as a master vehicle and all other vehicles in that region behave as slaves. All vehicles participated in the election and the master vehicle will be elected on the basis of its region time, speed and its good previous

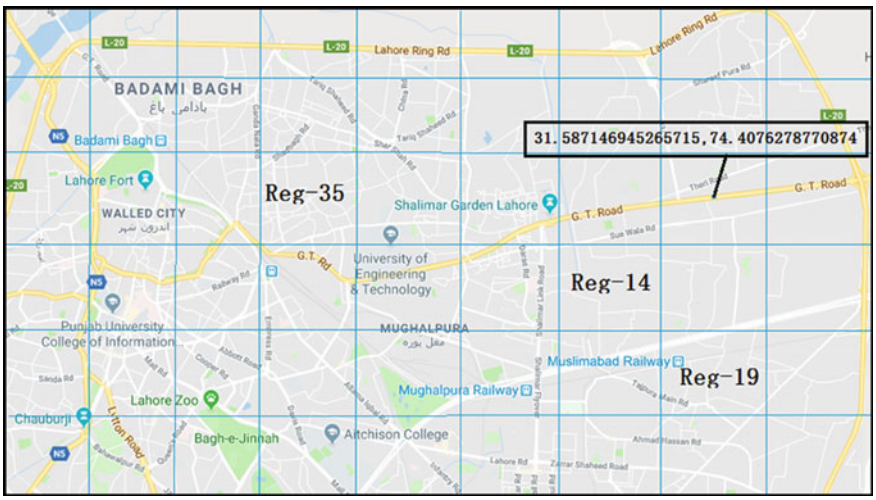


Fig. 1 Region based division

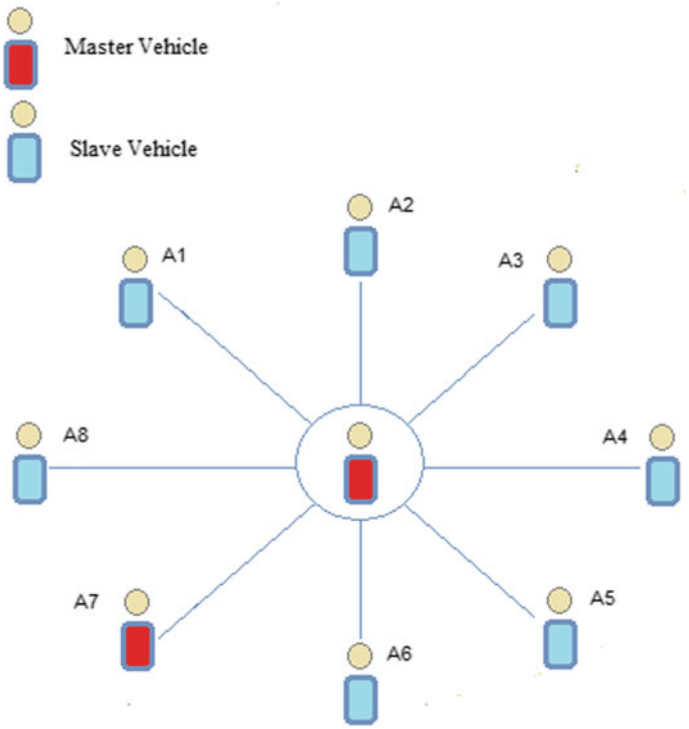


Fig. 2 Region topology

record (not involved in any malicious activity). The vehicles having the same region identity (Reg-Id) can communicate with each other. As the master vehicle enters in another region or its tenure period ends the election will be held and a new master vehicle will be elected again in that region. Each Region has only one master vehicle and many slave vehicles as shown in Fig. 2. The active smart vehicles A1, A2, A3, A4, A5, A6, A7, and A8 participate in the election to become a master vehicle. A7 wins the election and elected as a master vehicle. The master vehicle helps to load balance the social network computation. The master vehicle generates Sk_R and shared with all the slave vehicles for authentication of the broadcasted message.

Social Network creates a group of smart vehicles based on the region identity as each vehicle send its region identity to social network computed in Algorithm 1 using Haversine Formula [21]. So the vehicle can communicate with other vehicles in a region. Figure 3 shows a proposed architecture diagram for VANET.

Vehicle first get their coordinates (Lat_V, Lon_V) and then used haversine formula to calculate the nearest region using centered cordinates of all the regions as shown in Algorithm 1. In this paper we used the following notations (Table 1).

Fig. 3 Proposed architecture

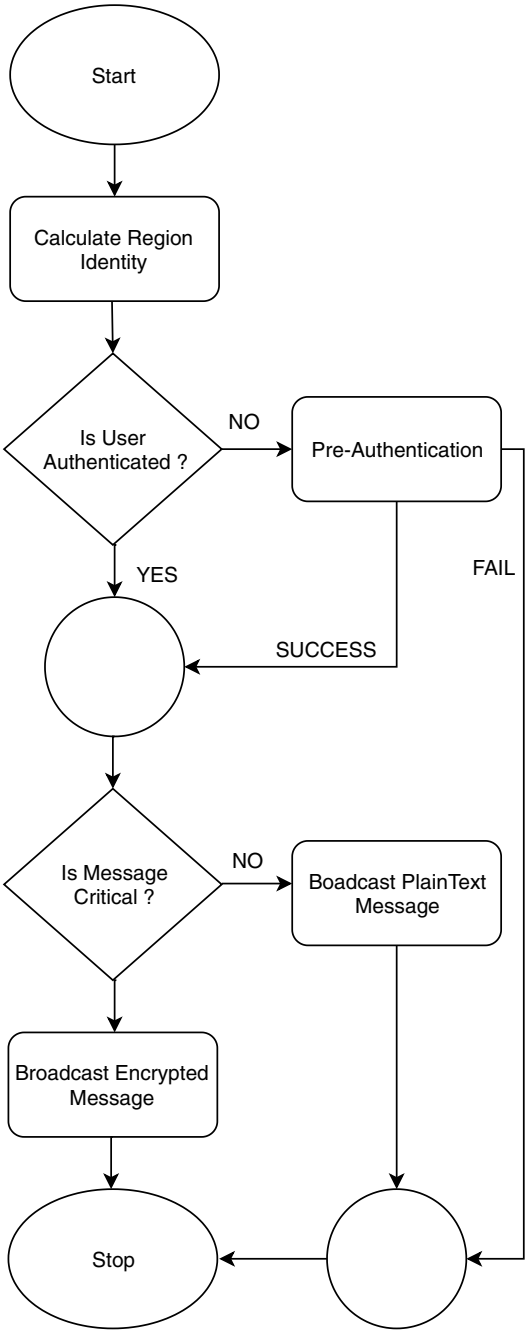


Table 1 Mathematical notations

Notation	Description
V	The vehicle
Reg-Id	Region identity
V_{ID}	Real identity of the vehicle
V_{RID}	Region identity of the vehicle
Sk_R	Secret key of region
Uid_M	User identity of Master vehicle
Uid_V	User identity of the vehicle
Pk_M	Public key of the master vehicle
Pk_{SN}	Public key of the social network
Prk_M	Private key of the master vehicle
Pk_V	Public key of the vehicle
Prk_V	Private key of the vehicle
Lat_C	Latitude of the center point of the region
Lon_C	Longitude of the center point of the region
Lat_V	Latitude of the vehicle
Lon_V	Latitude of the vehicle
R	Radius of the region
Msg	The message
$TmSt$	Timestamp
$Psid_V$	Pseudo Identity of the vehicle
n	Nonce

Algorithm 1 Region Identification

Require: Lat_V, Lon_V, Lat_C, Lon_C, Reg-Id, R
$r = 6373$: radius of the earth
dLat = Lat_C - Lat_V
dLon = Lon_C - Lon_V
$x = (sin(dLat/2))^2 + cos(Lat_C) * cos(LatV) * (sin(dLon/2))^2$
$y = 2 * atan2(sqrt(X), sqrt(1 - X))$
$d = r * y$: distance in Km
if $d <$ (R+ constant) **then**
V_{RID}= Reg-Id
end if

3.3 *Offline Registration*

During the offline registration, V computes its V_{RID} as shown in Algorithm 1, generates a pair of ECC Pk_V/Prk_V, and send Pk_V, V_{RID} and V_{ID} to social network

Table 2 User identity

User identity	Data
…	…
Uid_V	$(V_{ID})_{PK(RA)} \mid\mid Pk_V \mid\mid V_{RID}$
…	…

Table 3 Pseudo identity

Pseudo identity	Data
…	…
$Psid_V$	$Uid_V \mid\mid Pk'_V$
…	…

via some secure channel(for example vehicle visits Social Network). It is performed only once.

Step 1: V $\rightarrow$ SN $V_{ID}||Pk_V||V_{RID}$

Social Network verifies the V_{ID}, encrypt it with one of the public key generated by RA and store the parameters against the generated unique User identity Uid_V in the database as shown in the Table 2.

Social network encrypts the Uid_V, Uid_M, and Pk_M of the region with the Pk_V and send it back to the vehicle. The receiving vehicle will decrypt the message with its private key Prk_V and save the parameters in its secure OBU.

Step 2: SN $\rightarrow$ V $E(Pk_V, Uid_V||Uid_M||Pk_M)$

3.4 Pre-authentication

During pre-authentication, vehicles generate new ECC Pk'_V/Prk'_V and encrypt the Pk'_V, V_{RID}, and Uid_V with Pk_M and send it to V_M.

Step 3: V $\rightarrow$ V_M $E(Pk_M, Uid_V||V_{RID}||Pk'_V)$

Master vehicle decrypts the received message with its private key Prk_M and verifies the V_{RID} and save the message in its database against a unique Pseudo identity as shown in the Table 3.

Master vehicle encrypts generated Sk_R and $Psid_V$ with Pk'_V and sends it back to the vehicle.

Step 4: V_M $\rightarrow$ V $E(Pk'_V, Sk_R||Psid_V)$

If the master vehicle moves to another region, the election will held again and a new master vehicle will be elected in a region and generates new Sk'_R. When the slave vehicles or master vehicle moves to another region, slave vehicles compute its

V'_{RID} again, generates new ECC Pk''_V/Prk''_V and encrypt V'_{RID}, Pk''_V with Pk_{SN} and send it to the social network.

Step 5: V → SN $E(Pk_{SN}, V'_{RID}||Pk''_V||Pk_V||Uid_V)$

Social network decrypts the received message with its private key, verify it by matching Uid_V and Pk_V with the database parameters. If matched then update V'_{RID} and Pk''_V of the vehicle in the database and encrypt Uid'_M and Pk_M of the region with Pk''_V and send it back to the vehicle. The receiving vehicle decrypts the message with its private key.

Step 6: SN → V $E(Pk''_V, Uid_M||Pk_M)$

3.5 The Broadcsted Message

Vehicles prepare Msg by generating new ECC Pk'''_V/Prk'''_V and sign the message with Prk'''_V, encrypt the associated public key Pk'''_V with Sk_R along with V_{RID} and $Psid_V$.

Step 7: Msg = Sign (message,Prk'''_V) || E (Sk_R, Pk'''_V) ||V_{RID}|| $Psid_V$

The vehicles having the same V_{RID} will receive the message and decrypt the signature with the associated public key.

4 Analysis and Results

This section presents the security analysis and performance evaluation of our proposed protocol. Security analysis satisfies the basic authentication properties and performance evaluation shows the efficiency of our proposed protocol.

4.1 Security Analysis

In this section, we analyze the security analysis of our proposed protocol.

- Message Integrity: The receiving vehicles ensure the integrity of the received Msg by verifying the signature with the associated public key.
- Vehicle Authentication: Social network authenticates the vehicle by decrypting the received message with its private key and compare the Uid_V and Pk_V with the database values.
- Privacy Preservation: In our protocol, RA has a private key to decrypt the real identity of the vehicle (V_{ID}) but has no access to the social network database. Even if the database server is compromised the real identity of the vehicle will not

Table 4 Parameters

Parameters	Value
Message size	324B
Number of regions	100
Network	LTE
Vehicle speed	15–25 m/s

be leaked. Whereas vehicle used it's $Psid_V$ and Uid_V to communicate with other vehicles in a region.

- Non-repudiation: The sending vehicle prepares the Msg by appending the associated public key with the signature. The message also includes the nonce and current time stamp. Therefore each Msg is unique itself and it also prevents the replay attack.
- Vehicle Revocation: If the vehicle is involved in any malicious activity the master vehicle informs the social network about its behavior and the social network revoked the vehicle from its database. Master vehicle generates new Sk_R and sends it to all other vehicles in a region except the revoked vehicle. Now the malicious vehicle will not able to take part in the network. Even if the master vehicle is compromised the privacy of the vehicles still remained secure and the master vehicle is revoked after its tenure period ends.

4.2 Simulation Environment

The simulation of our proposed protocol is carried out using network simulator ns-3. The size of the safety information broadcasted is considered as: Message = 200B standard size, signature = 64B, encryption = 48B, Pseudo identity = 6B and region-Id = 6B. So the total communication overhead bytes are 124 which shows that our proposed protocol is lightweight. We considered the following parameters for simulating our proposed protocol (Table 4).

4.3 Performance Evaluation

We evaluate the performance of our proposed protocol by comparing with Rajput protocol [14]. Our testbed includes Intel i3 processor with 6 GB of RAM and the computation is carried in java container as java supports a rich set of cryptographic

Table 5 Performance

	Overhead	Time (ms)
Vehicle speed	Region calculation	0.0038
	ECC Pk/Prk key generation	2
	ECC signature generation	1.884
	AES encryption	0.357
Receiving vehicle	ECC signature verification	2.946
	AES decryption	0.253
Master vehicle	AES key generation	0.01289
	ECC encryption	9.267
	ECC decryption	7.814

libraries. Our proposed protocol is more efficient and light-weight besides the fact that Rajput [14] testbed is based on Intel i7 processor with 16 GB of RAM. The communication overhead bytes of our proposed protocol are 124 whereas the communication overhead bytes of Rajput protocol are 162 (Table 5).

- Computation Overhead: Computation overhead is the overall time taken by the receiving vehicles to verify the broadcasted message. Receiving vehicle first decrypt the associated Pk of the signature with the Sk and then decrypt the signature. Our proposed protocol takes around 3.2 ms whereas the Rajput protocol [14] takes 5 ms as shown in the figure. For encryption/decryption, we used AES with PKCS5 padding and CBC mode of operation using random initial vector (IV) and key size is 128 bits. The signature is computed using ECDSA with PKCS8 padding and SHA256 digest. We run the test 100 times to take the average results shown in the table.
- Packet Loss: Packet loss is one of the most important metrics to evaluate network performance. We observed during the simulation that the packet loss increases as distance increases from the access point as shown in the figure. However, the packet loss difference between secure and unsecured message is around 6 dB with the least distance from the access point and it decreases as the distance increases from the access point (Figs. 4 and 5).

5 Conclusion

This paper proposes a dynamic efficient privacy-preserving authentication protocol in VANET using the social network. We propose the novel idea of dynamic region topology which helps to ensure that the message gets delivered to desired vehicles in a region using the social network. The master vehicle helps to load balance the social network server computation. We adopt a hybrid combination of symmetric AES

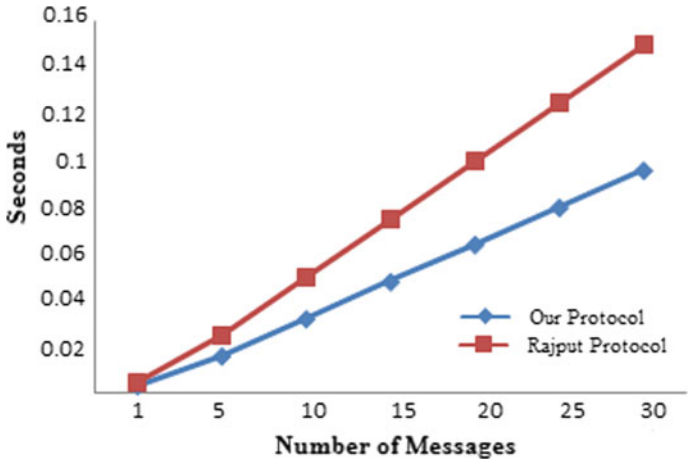


Fig. 4 Compound overhead

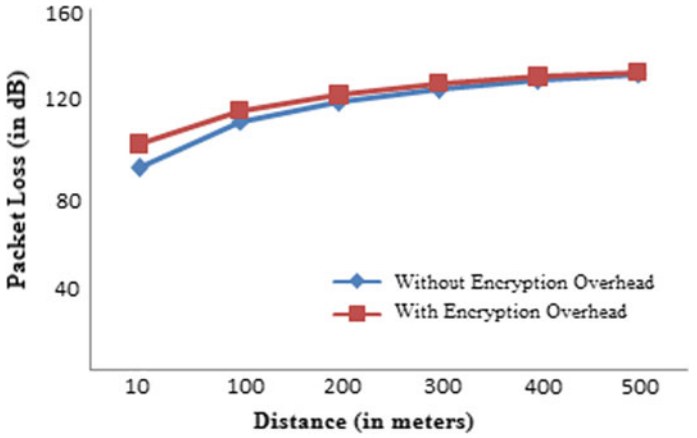


Fig. 5 Packet loss

and asymmetric ECC based cryptography for efficient authentication. The security analysis and simulation results show that our protocol is secure, privacy-preserving, light-weight and provide better performance by reducing the communication and computation overhead.

Acknowledgements The authors are grateful for insightful and helpful conversations with Dr. Hongsheng Zhou, and Renjun Zhang and Zengpeng Li. The authors would also like to thank the anonymous referees for their helpful suggestions and valuable comments. The work is supported by the following grants: National Key Research and Development Program of China (2017YFB0802500, 2017YFB0802505, 2017YFB0802203, 2018YFB1003701) and National Natural Science Foundation of China (U1736203, 61732021, 61472165, 61373158) and the China Scholarship Council.

References

1. Road safety annual report. https://www.itf-oecd.org/road-safety-annual-report-2018. Accessed 18 May 2018
2. Mallissery, S., Pai, M.M.M., Pai, R.M., Smitha, A.: Cloud enabled secure communication in vehicular ad-hoc networks. In: 2014 International Conference on Connected Vehicles and Expo, ICCVE 2014 - Proceedings, pp. 596–601 (2014)
3. Lin, X., Lu, R.: Vehicular Ad Hoc Network Security and Privacy. Wiley, New York (2015)
4. Tangade, S.S., Manvi, S.S.: A survey on attacks, security and trust management solutions in VANETs. In: 2013 4th International Conference on Computing, Communications and Networking Technologies ICCCNT 2013, pp. 1–6 (2013)
5. Lin, X., Li, X.: Achieving efficient cooperative message authentication in vehicular ad hoc networks. IEEE Trans. Veh. Technol. **62**(7), 3339–3348 (2013)
6. Zhang, C., Lu, R., Lin, X., Ho, P.H., Shen, X.: An efficient identity-based batch verification scheme for vehicular sensor networks. In: Proceedings - IEEE INFOCOM, pp. 246–250 (2008)
7. Engoulou, R.G., Bellache, M., Pierre, S., Quintero, A.: VANET security surveys. Comput. Commun. **44**, 1–13 (2014)
8. de Fuentes, J.M., Gonzalez-Tablas, A.I., Ribagorda, A.: Overview of security issues in vehicular ad hoc networks. In: Cruz-Cunha, M.M., Moreira, F. (eds.) Handbook of Reseach on Mobility and Computing, pp. 1–17. IGI Global, Pennsylvania (2010)
9. Raya, M., Hubaux, J.: The security of vehicular ad hoc networks. In: Proceedings of the 3rd ACM workshop on Security of Ad hoc and Sensor Networks, November 2005, pp. 11–21 (2005)
10. Varshney, N., Roy, T., Chaudhary, N.: Security protocol for VANET by using digital certification to provide security with low bandwidth. In: International Conference on Communications and Signal Process, ICCSP 2014 - Proceedings, pp. 768–772 (2014)
11. Tangade, S.: Scalable and privacy-preserving authentication protocol for secure vehicular communications. In: IEEE International Conference on Advanced Networks and Telecommunications Systems ANTS 2016, pp. 1–6 (2016)
12. Sun, Y., et al.: An efficient pseudonymous authentication scheme with strong privacy preservation for vehicular communications **59**(7), 3589–3603 (2010)
13. Rhim, W.W.: A study on MAC-based efficient message authentication scheme for VANET. M. S. thesis, Hanyang University (2012)
14. Rajput, U., Abbas, F., Oh, H.: A hierarchical privacy preserving pseudonymous authentication protocol for VANET. IEEE Access **4**, 7770–7784 (2016)
15. Lu, R., Lin, X., Zhu, H., Ho, P., Shen, X.S.: ECPP: efficient conditional privacy preservation protocol for secure vehicular communications, pp. 1229–1237 (2008)
16. Li, J., Member, S., Lu, H., Guizani, M.: ACPN: a novel authentication framework with conditional privacy-preservation and non-repudiation for VANETs. IEEE Trans. Parallel Distrib. Syst. **26**(4), 938–948 (2015)
17. Zhang, L., Hu, C., Wu, Q.: Privacy-preserving vehicular communication authentication with hierarchical aggregation and fast response. IEEE Trans. Comput. **65**(8), 2562–2574 (2016)
18. Shamir, A.: Identity-Based Cryptosystems and Signature Schemes. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 196 LNCS, pp. 47–53 (1985)
19. Jiang, S., Zhu, X., Wang, L.: A conditional privacy scheme based on anonymized batch authentication in Vehicular Ad Hoc Networks. In: 2013 IEEE Wireless Communications and Networking Conference, pp. 2375–2380 (2013)
20. Paranjothi, A., Khan, M.S., Nijim, M., Challoo, R., MAvanet: message authentication in VANET using social networks. In: IEEE 7th Annual Ubiquitous Computing. Electronics and Mobile Communication Conference, UEMCON 2016, pp. 1–8 (2016)
21. Haversine Formula. https://andrew.hedges.name/experiments/haversine/

A Negotiation Strategy Based on Compromising Degree

Shun Okuhara and Takayuki Ito

Abstract This paper proposed a compromising strategy based on constraint relaxation for automated negotiating agents. A lot of studies including international competitions have been made on automated negotiating agents. basically, automated negotiating agents are adjusting a threshold to accept their opponents' offers. Because merely a threshold is adjusted, it is very difficult to show how and what the agent conceded even after an agreement has been reached. To address this issue, we propose an explainable concession process using a constraint relaxation process. Authors also describe methods min, distance and distance min, random for the size of the removed constraint. Experimental results demonstrate that distance strategies are effective.

1 Introduction

Multi-agent systems are one of the most important technological advancements that have been made to address the needs of the next generation [1–8]. The ANAC (Automated Negotiating Agents Competition) has been held since 2010 as a testbed for automatic negotiation agent research [9]. ANAC adopts a multi-issue utility model and an alternating-offer protocol, and changes and extends the rules of negotiations every year. However, there are several drawbacks and problems that the ANAC competition could not focus on. One of them is how to explain the compromise process. In negotiations, agents cannot reach an agreement if they consider only their own profits and interests. Therefore, the compromise strategy is essential to reach an agreement. Most of the existing automated negotiating agents adopt ad hoc compromising pro-

S. Okuhara (✉)
School of Medical Sciences, Fujita Health University, 1-98 Dengakugakubo, Kutsukake-Cho, Toyoake, Aichi, Japan
e-mail: okuhara@itolab.nitech.ac.jp

T. Ito
Graduate School of Engineering, Nagoya Institute of Technology, Gokiso-Cho, Showa-Ku, Nagoya, Aichi, Japan
e-mail: ito.takayuki@nitech.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_5

cesses that only adjust their thresholds to accept the opponent's offer. This has made it difficult to explain how the compromise was achieved in the negotiation. To solve this problem, we propose a compromise process based on constraint relaxation. A constraint is a basic unit of utility. In other words, the utility space of an agent is regarded as a set of constraints that satisfy the issue values and the argument for them. When a constraint is satisfied, the agent gets a utility value for this constraint. We also assume shared issues and individual issues, which means that agents agree if they have the same issue value for shared issues. For individual issues, each agent can choose issue values to make their utility as high as possible. An agent faces a tradeoff between maximizing its own utility by satisfying the constraint as much as possible, while keeping the share value to be the same value as that of the opponent agent. In order to settle this tradeoff, agents perform compromising. In the compromise process for the strategy we propose, the agent removes constraints one by one from the set of its own constraints. Then, it tries to change the constraints' most preferable issue-value of the shared issue. If the agent can change the issue-value to one that is the same as the opponent's, then they can reach an agreement. Removing constraints is called constraint relaxation. Specifically, we assume that the agent has a believed constraint set (IN) and an unbelieved (OUT) constraint set. In the initial state, it is assumed that all constraints are IN, and that in the constraint relaxation process, agents move certain constraints from IN to OUT. Various strategies are enabled when the agent moves constraints from IN to OUT. The four methods we propose are:

(1) Relaxation of constraints based on value (hereinafter called "Min"),
(2) Random constraint relaxation (hereinafter called "Random"),
(3) Constraint relaxation based on distance (hereinafter called "Distance"),
(4) Constraint relaxation based on value and distance (hereinafter called "Distance min").

Authors obtained demonstrate that methods (3) are able to obtain social surpluses significantly higher than the (2) random constraint relaxation at Experimental results. In Sect. 2, we describe the automatic negotiation agent and negotiation protocol and then, a compromised algorithm that is based on the newly proposed constraint relaxation in Sect. 3. In Sect. 4, we describe and discuss experimental results. In Sect. 5, we clarify the difference between our methods and related research work. Finally, we summarize our paper in Sect. 6.

2 Automated Negotiating Agents

2.1 Utility Hyper-graph

An agent possesses a complex utility space [10]. Such complex utility spaces have been presented in many different ways [11–13]. In this paper, our approach is to represent the way with hypergraph, to focus on the mutual dependence between

issues(nodes). A hypergraph is one of the mathematical representations whose edge can combine multiple nodes. A utility using hypergraph is called utility hyper-graph. Regarding this representation, we consider nodes as issues, also consider edges as constraints. Hypergraph (I, C) represents the utility space of agent I_i I. In the utility space, I mean a set of issues (nodes), and C means a set of constraint (edge). Each I_i I issue possesses issue value (Issue Value) in a certain range D_i. For example, color, i.e., one of the issues when purchasing a car possesses one issue value in the following rangered, blue, green.

$(v_{C_j}, \phi_{C_j}, \delta_{C_j})$ represents Constraint C_j C. v_{C_j} and ϕ_{C_j} is a set of issues in which constraints (i.e. C_j) are combined.

Accordingly, Consequently, δ_{C_j} is a set of ranges where $\delta_{C_j} = \{range_{C_j}(I_i) : I_i \in C_j\}$. The conditions under which constraint C_j is satisfied are as follows. The value assumed by issues I_i is x_{I_i}. If C_j is satisfied, then an agent having C_j obtains the value thereof v_{C_j}.

Figure 1 shows an example of an agent's utility graph and issues shared.

Here, two agents who have their own utility graph share three issues. Each of the agents has constraints that link issues. The issue takes an issue value. A constraint is satisfied if the issues linked by this constraint have issue values within the predefined ranges. When a constraint is satisfied, the agent obtains a value from this satisfied constraint. An example sharing utility graph of agent and issues is shown in Fig. 1. It shows that two agents have their own utility graph and three issues simultaneously. There are constraints that link each issue in the respective agents. An issue takes its value. The value meets a constraint if the value of the issue linked with the constraint is included in the constraint's range. When the value meets the constraint, the agent will be able to get a value from the constraint. Assumption 1: The more difficult to satisfy a constraint is, the higher it's value becomes. We made 2 assumptions according to assumption 1 as follows. Value of a constraint which has wider issue-

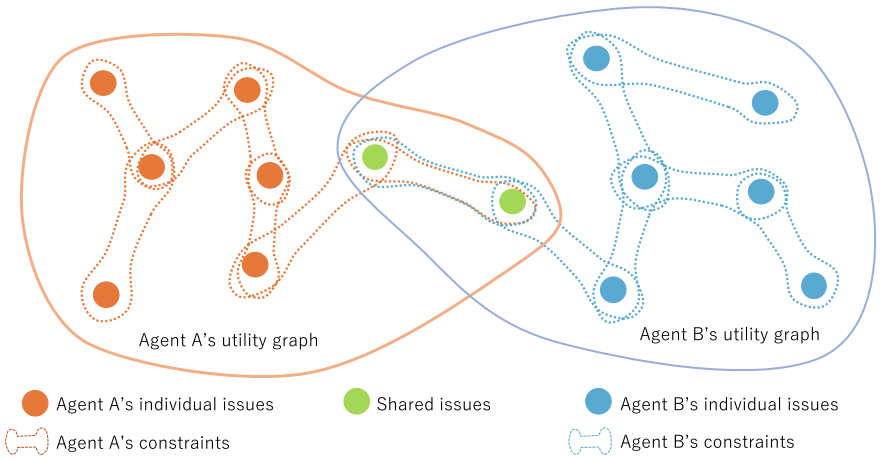


Fig. 1 Sharing issue and utility graph

value $range_{C_j}(I_i)$ is lower because it is easier to satisfy. Meanwhile, value of a constraint which has narrower issue-value range is higher because it is More difficult to satisfy. Moreover, a shared constraint which needs agreement with an opponent has higher value than individual constraint.

2.2 Negotiation Protocol

At this time, we use negotiation protocol which is as simple as possible for the purpose of focusing on the compromise algorithms only. Our proposal is a simultaneous repeated offer protocol, that is, each agent submits a proposal to the opponent at the same time and get mutual agreement if both can accept the proposal. If neither of them accepted, both agents submit the revised proposal again by compromising. This proposal would be repeated till when one of the agents gives up on compromising. The specific algorithm is as follows:

1: Each of the agents generates the most suitable proposal for itself (the most suitable assignment for each issue). 2: Each of the agents proposes a value for the shared issue at the same time. 3: Judging agreement. 4: If the value offered by both agents for shared issues are the same. 5: The protocol is ended by an agreement between them. 6: Or by reaching predetermined numbers of the proposal. Back to 1. if the result is in all other cases. 7: Each agent performs the compromise process (refer to the next section). 8: end if 9: until one of the agents cannot continue, i.e., no constraint can be relaxed or when the prescribed number of iterations is reached.

An agent modifies its utility space by doing a compromise process and makes the most suitable proposal to its opponent. That is, each agent makes the most suitable proposal based on a utility space of its own. Mutual proposal protocol has also been used in the research field of automated negotiations, however, we used simple simultaneous repeated offer protocol, otherwise, a compromise strategy would be changed. It is further working to extend simultaneous repeated offer protocol to Mutual proposal protocol.

3 Explainable Compromise Process Based on Constraint Relaxation

3.1 Explainable Compromise Process

This section describes compromise process based on constraint relaxation. Simple examples of compromise process which is proposed in this section are shown in Figs. 2, 3, 4, 5. We define constraint relaxation as reducing the sum of utilities (value) that one supposes to obtain by reducing the number of constraints which must be satisfied. Existing research which makes a compromise by ad hoc thresh-

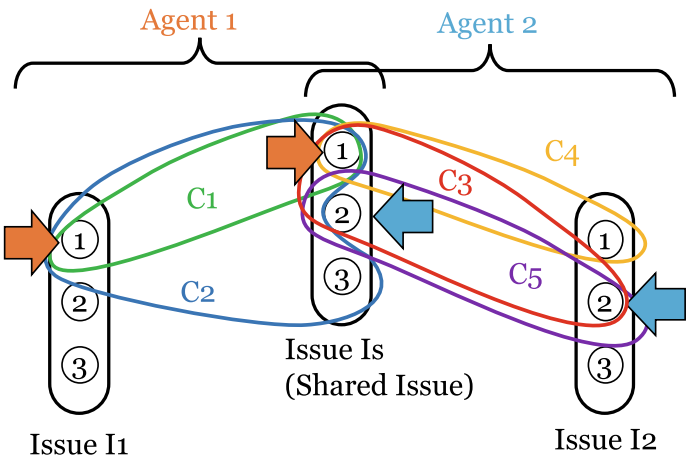


Fig. 2 Case 1: initialization

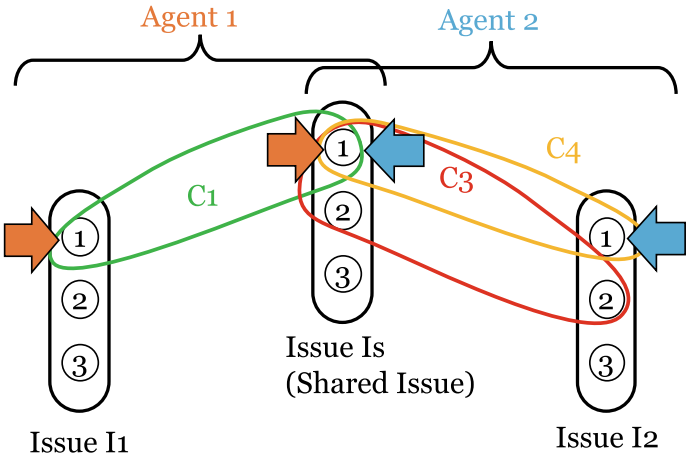


Fig. 3 Case 2: agreement by relaxation

old adjustment is unable to explain how it leads to agreement with its value. We reduce constraints which must be satisfied in this research. Specifically, we don't take constraints into consideration so that it enables us to explain of compromise, such as which constraint is taken in consideration and which is not when reaching an agreement. Figures 2 and 3 show a case where there is an agreement.

Figure 2 shows Agent1 possesses Issue I1 and Issue Is. Is is a shared issue. Agent 2 possesses Issue I2 and Issue Is. Each of the issues possesses 3 values, i.e., 1, 2, and 3. Agent1 possesses constraint C1 and C2. Regarding Issue I1, the initial optimal solution is 1 and regarding Is, the initial optimal solution is 1. It is because when both issues are satisfied, the utility is higher. Meanwhile, Agent 2 possesses constraint C3,

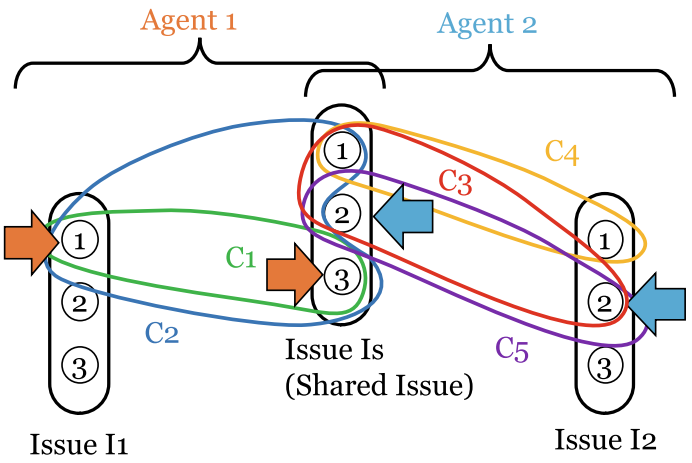


Fig. 4 Case 3: initialization

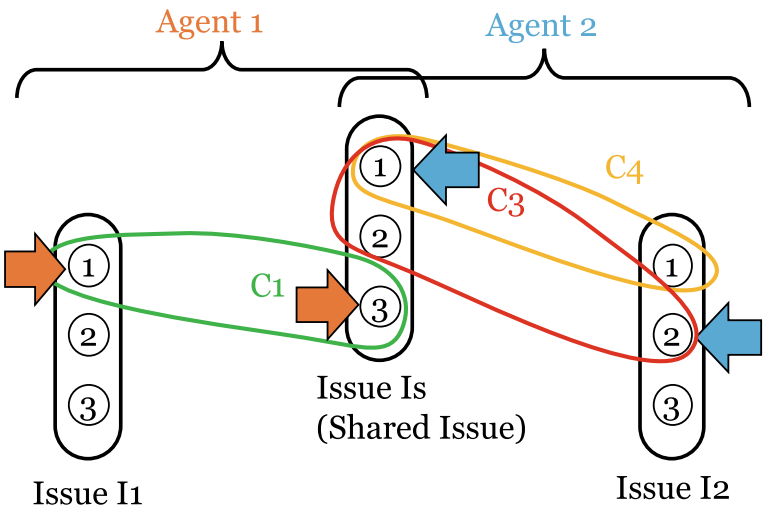


Fig. 5 Case 4: agent cannot an agreement

C4 and C5. Similarly, regarding Is, optimal solution is 2, and regarding I2, optimal solution is 2. There is no agreement in the condition of Fig. 3 because the solution is different regarding shared issue. Hence, each agent removes (change from IN to OUT) one constraint and make a compromise process. For instance, constraint C2 is set as OUT in Agent1, and constraint C5 is set as OUT in Agent 2 in this case. When that happens, the value of Issue Is in Agent 1 stays at 1. On the other hand, the value of Issue Is in Agent 2 also becomes 1. The result enables both Agent1 and Agent2 to reach the agreement. As for compromise, it enables us to know which constraint

is set as OUT (not to believe). Therefore, it allows us to explain which constraint is excluded, not only reducing threshold.

On the other hand, Figs. 4 and 5 show a case where there is not an agreement.

For instance, constraint C2 is set as OUT in Agent1, and constraint C5 is set as OUT in Agent 2. In this case, agents cannot reach an agreement. the above cases are just examples. However, because of the value, the result of negotiation would change very much.

3.2 Compromising Strategies

Although various relaxed constraints are considered, we propose 4 strategies as follows in this section. We set all the initial constraints as IN, and all the relaxed constraints as OUT.

- Relaxed constraint based on Randomness: a constraint which is randomly selected among IN is changed into OUT.
- Relaxed constraint based on value: a constraint with the lowest value among IN is selected and changed into OUT.
- Relaxed constraint based on distance*: a constraint which is the farthest from the shared issue among IN is selected and changed into OUT. *At this point, distance is defined as the number of constraints which are connected from the shared issue.
- Relaxed constraint based on min distance: a constraint which is the farthest from the shared issue and with the lowest value among IN is selected and changed into OUT.

4 Experiment

4.1 Experiment Setting

We performed an experiment to compare the performances of the proposed compromising strategies. Our experimental setting included the following parameters:

- Agents are two.
- One issue can include up to 10 values.
- One issue is a shared issue.
- Each issue has an issue at least one constraint.

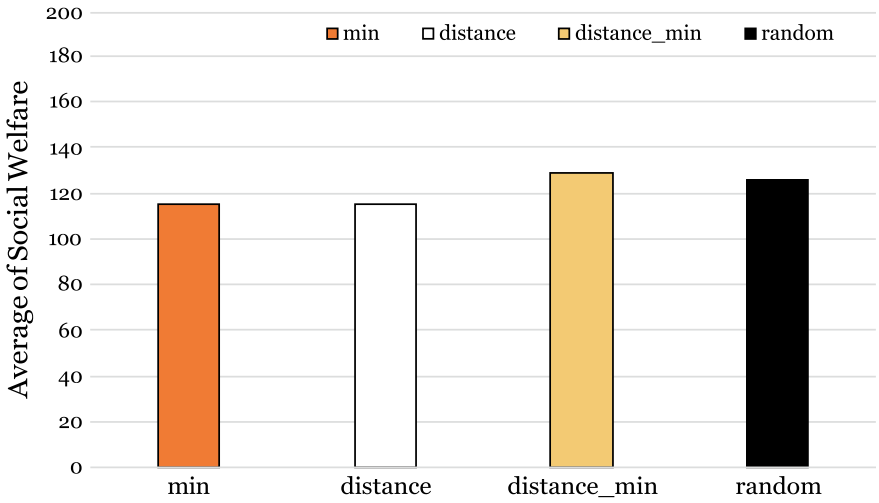
In other words, each issue is consists of one or more constraints.

Authors employed a multi-start local search approach as a search method to find the optimal solution. Issues are assigned randomly, and Graph structures are based on constraints.

Our experimental setting implies a situation where all of them are connected with a number of constraints. Authors ran 1000 trials. In one trial, agents iterate at most N offers with the proposed simultaneous repeated offer protocol where N is the number of constraints. The maximum number of iteration is the number of constraints because an agent removes a single constraint for each iteration. In a single iteration, each agent optimizes their issue-values by using the multiple local searches. Where we set the value of 100 different restarts and 100 steps to search for each search.

4.2 Results and Discussion

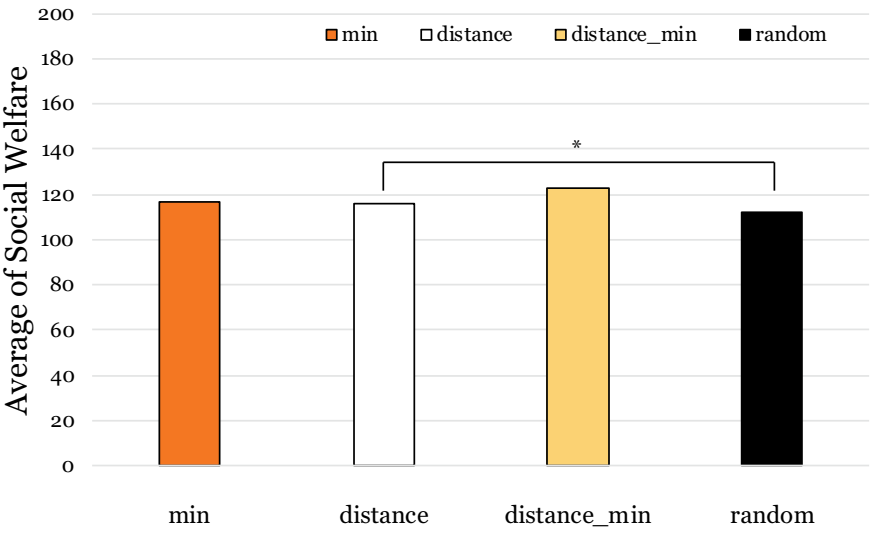
This paper presents the results of several settings. The results obtained four settings which are shown in Figs. 6, 7, 8, 9. These figures compare the social welfare in graph.



	the number
Number of Constraints Per Issue	1
Number of Issue Per Agent	5
Number of Constraints	5
Max Constraint Dimensions	3
Concession Degree	1

comparisons for all pairs using Tukey-Kramer HSD
*:P <0.05, statistically significant **:P <0.01, statistically highly significant

Fig. 6 Experimental results for $x = 1$



	the number
Number of Constraints Per Issue	1
Number of Issue Per Agent	5
Number of Constraints	5
Max Constraint Dimensions	3
Concession Degree	2

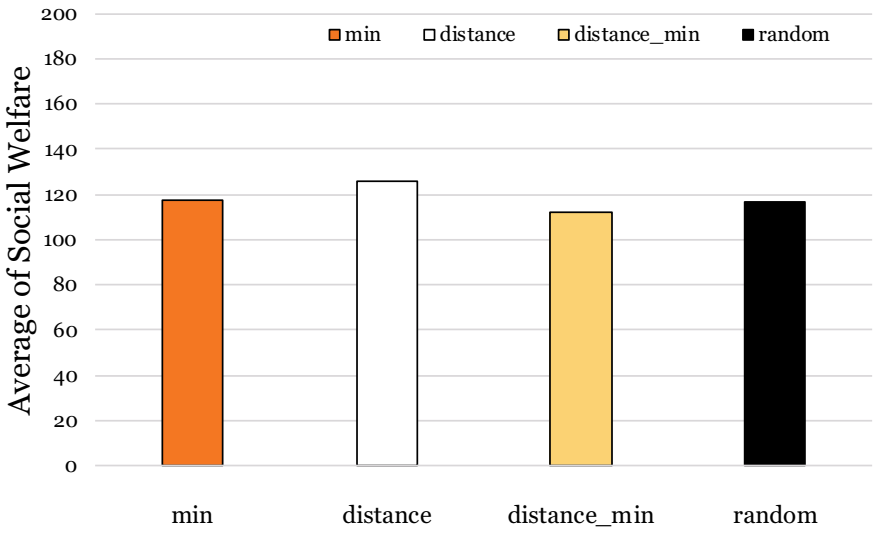
comparisons for all pairs using Tukey-Kramer HSD
*:P <0.05, statistically significant **:P <0.01, statistically highly significant

Fig. 7 Experimental results for $x = 2$

Authors use the Tukey's honest significance test for comparisons for all pairs of the samples. The Figures shows the number of samples at n. x is the number of removed constraints. Figure 6 shows the experimental results for $x = 1$. Distance min got higher social welfare compared at Random. It is not statistically significant in this experiment. There is no significant difference between Min, Distance.

Figure 7 shows the experimental results for $x = 2$. Distance got higher social welfare compared with Random. It is statistically significant. Distance strategy works better than Random strategy in this case.

Figure 8 shows the experimental results for $x = 3$. Distance got higher social welfare compared with Min and Distance min and Random. It is not statistically significant in this case.



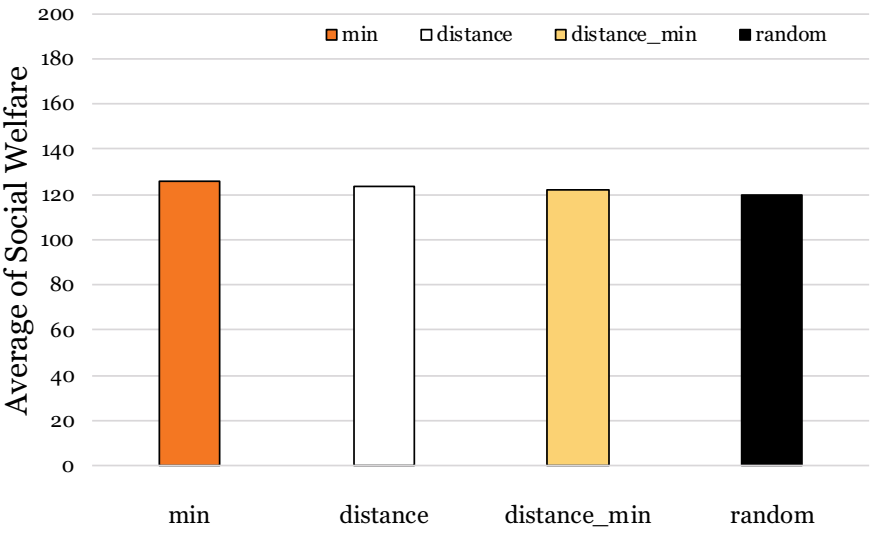
	the number
Number of Constraints Per Issue	1
Number of Issue Per Agent	5
Number of Constraints	5
Max Constraint Dimensions	3
Concession Degree	3

comparisons for all pairs using Tukey-Kramer HSD
*:P <0.05, statistically significant **:P <0.01, statistically highly significant

Fig. 8 Experimental results for $x = 3$

Figure 9 shows the experimental results for $x = 4$. Min, Distance, and Distance min Random did not show any differences regarding got not difference social welfare. It is not statistically significant.

When there were more than 3 removed constraint per agent, the authors were unable to get stable experiment results. Namely, it was difficult to obtain results showing a significant difference in the drawing method. This is because there are more than 3 removed constraint per agent. The number of solutions was less. The graph structure currently given to the agent is randomly given in small size. Developing an optimization strategy based on the structure of the graph will be also a subject for future work.



	the number
Number of Constraints Per Issue	1
Number of Issue Per Agent	5
Number of Constraints	5
Max Constraint Dimensions	3
Concession Degree	4

comparisons for all pairs using Tukey-Kramer HSD
*:P <0.05, statistically significant **:P <0.01, statistically highly significant

Fig. 9 Experimental results for $x = 4$

5 Related Work

In this section, we describe the differences between our study and related work. In the field of automated negotiation research, the compromise process was first proposed by Klein et al. [14]. His main argument is that it is reasonable for the agent to gradually compromise at the Pareto front in simple negotiations where the issues are independent, and the utility space is liner in each issue. However, if the issues are interdependent, the process is not simple because utility space is complicated, which makes the agent unable to find the Pareto front easily. To address this problem, Klein et al. proposed a SA-based agreement point search protocol (implicitly assuming compromising). In addition, Faratin et al. [15] analyzed various compromise functions.

The ANAC Competition [9] has been held annually since 2010. It is common for ANAC agents to adopt a method for estimating and presenting proposals that can be statistically accepted from the opponent's offers and accepting the proposal by adjusting the threshold considering the time discount utility. For example, AgentK [16], the winning agent of ANAC 2010, estimates the opponent utility space and the attitude (hostile or compromising) towards agreement from the opponent's offer history. If the partner seems to be a compromise, the concession is made, and if the partner is hostile, it will not concede more than a certain threshold. The above is the strategy that pioneered ANAC's basic concession strategy. Fawkes [9], the winning agent of ANAC 2013, estimates optimal concessions using discrete wavelet prediction based on an opponent's offer history. Most existing studies have focused on how to adjust the threshold so that the opponent's offer can be accepted. The threshold is a kind of upper limitation utility with which the agent can accept the opponent's offer. However, these studies give no explanation about how to archive the threshold value. Thus, they do not explain why the agent compromises. This is a real problem because if your self-driving car compromises, you will not be able to obtain any explanation about the compromise. Also, as far as the authors know, no research has been done that assesses how compromising can be explained in an automated negotiation agent that assumes multi-argument utility functions.

Sycara has published a series of studies [17, 18] that proposed negotiation and compromising processes that are explainable because they use case-based reasoning. The point is that they defined compromise and persuasion in the form of logical arguments within the framework of case-based reasoning. Sycara's series of studies is also related to argumentation theory [19, 20] and has developed into mathematical argumentation theory. On the other hand, the viewpoint of this research is focusing on how to construct an explainable compromise process based on the utility function that can be handled numerically.

The Distributed Constraint Satisfaction/Optimization Problem (DCSP/DCOP) [21, 22] has been one of the major topics in multi agent research. Because our model is based on constraints, it is closely related to DCSP/DCOP. The main difference is that our model focuses on negotiation situations where agents are basically trying to maximize their own individual utilities, but they compromise because they need to make an agreement. This is because if they cannot reach an agreement, there is no utility. In DCSP/DCOP, however, agents basically do not consider their own individual utilities. The main focus is on constraint satisfaction or optimization with distributed cooperative agents.

Wakaki et al. [20] published a paper about a DTMS (Distributed Truth Maintenance System) in which they proposed a classification of *consistency* in multi-agent environments. They classified the distributed consistency concept into Inconsistent, Local-Consistency, Local-and-Shared-Consistency, and Global Consistency categories. In this study, an agreement means that each agent has its internal consistency while they have a consistently shared issue-value, which is the Local-and-Shared-Consistency. The compromising method we propose is one of the methods for obtaining Local-and-Shared-Consistency. However, the constraint graph we use represents the utility space. On the other hand, a DTMS does not express preferences.

6 Conclusion

We have proposed an explainable compromise process for automatic negotiation. In previous studies, an agreement was reached when the utility value simply became greater than the threshold. Therefore, it was extremely difficult to explain how agreement among agents was arrived at in a way that was understandable by humans. To resolve this problem, the authors proposed a model capable of implementing a concurrence process by removing constraints and explaining the reason for the agreement. In this study, we investigated applying the min, distance, distance min, and random methods of constraint removal. The min method removes constraints in order of size from smallest to largest. The distance method removes constraints from beginning with those that have the farthest local issues. The distance min method first removes the constraints from shared issues that are both the fastest and smallest. The random method randomly selects constraint values to remove.

Upon comparing these four methods, this study confirmed that the distance method exhibited a higher social welfare value than the random method and also produced statistically significant values. However, this study was not able to confirm any difference in the social welfare values of each method in cases where three or more constraints had been removed.

The following are our contributions: (1) The novel explainable compromise process we developed is based on a utility graph-structured with constraints and issues. (2) For the compromise process, we developed a constraint relaxation process based on distance and value and demonstrated its effectiveness.

As a subject for future work, we will need to further expand the size of the problem space and perform an evaluation experiment to more closely investigate the effect of removing a given number of constraints.

Furthermore, we should attempt to develop a more sophisticated compromising process. For example, it should be possible to create a process that can find possible combinations of the fewest constraints to be relaxed so that agents can change their alternatives. Alternating Offer protocol [23] has been one of the popular protocol for research on bilateral negotiating agents for long years. In this paper, we employed a simple simultaneous offering protocol because our main claim in this paper is to propose a new compromising process for agents. It would be another subject for future work to extend our protocol to the alternating offers protocol.

References

1. Bai, Q., Ren, F., Fujita, K., Zhang, M., Ito, T.: Multi-agent and Complex Systems. Springer, Singapore (2017)
2. Fukuta, N., Ito, T., Zhang, M., Fujita, K., Robu, V.: Recent Advances in Agent-Based Complex Automated Negotiation. Springer, Switzerland (2016)
3. Fujita, K., Ito, T., Zhang, M., Robu, V.: Next Frontier in Agent-Based Complex Automated Negotiation. Springer, Japan (2015)

4. Marsa-Maestre, I., Lopez-Carmona, M.A., Ito, T., Zhang, M., Bai, Q., Fujita, K.: Novel insights in Agent-Based Complex Automated Negotiation. Springer, Japan (2014)
5. Ito, T., Zhang, M., Robu, V., Matsuo, T.: Complex Automated Negotiations: Theories, Models, and Software Competitions. Springer, Berlin (2013)
6. Ito, T., Zhang, M., Robu, V., Fatima, S., Matsuo, T.: New Trends in Agent-Based Complex Automated Negotiations. Springer, Berlin (2011)
7. Ito, T., Zhang, M., Robu, V., Fatima, S., Matsuo, T., Yamaki, H.: Innovations in Agent-Based Complex Automated Negotiations. Springer, Berlin (2010)
8. Ito, T., Zhang, M., Robu, V., Fatima, S., Matsuo, T.: Advances in Agent-Based Complex Automated Negotiations. Springer, Berlin (2009)
9. Baarslag, T., Fujita, K., Gerding, E., Hindriks, K., Ito, T., Jennings, N.R., Jonker, C., Kraus, S., Lin, R., Robu, V., Williams, C.: The first international automated negotiating agents competition. Artif. Intell. J. (2012)
10. Ito, T., Hattori, H., Klein, M.: Multi-issue negotiation protocol for agents: exploring non-linear utility spaces. In: Proceedings of 20th International Joint Conference on Artificial Intelligence, pp. 1347–1352 (2007)
11. Robu, V., Somefun, D.J.A., Poutr'e, J.L.: Modeling complex multi-issue negotiations using utility graphs. In: AAMAS 05: Proceedings of the Fourth International Joint Conference on Autonomous agents and Multiagent Systems. ACM, New York, NY, USA, pp. 280–287 (2005)
12. Robu, V., La Poutr'e, H.: Constructing the structure of utility graphs used in multi item negotiation through collaborative filtering of aggregate buyer preferences. In: Rational, Robust, and Secure Negotiations in Multi-Agent Systems, pp. 147–168. Springer, Berlin (2008)
13. Aydogan, R., Baarslag, T., Hindriks, K., Jonker, C., Yolum P.: Heuristics for using cp-nets in utility-based negotiation without knowing utilities. In: Knowledge and Information Systems, vol. 45, pp. 357–388, 11 (2015)
14. Klein, M., Faratin, P., Sayama, H., Bar-Yam, Y.: Negotiating complex contracts. Group Decis. Negot. **12**(2), 58–73 (2003)
15. Faratin, P., Sierra, C., Jennings, N.R.: Negotiation decision functions for autonomous agents. Int. J. Robot. Auton. Syst. **24**(3–4), 159–182. http://eprints.ecs.soton.ac.uk/2117/ (1998)
16. Kawaguchi, S., Fujita, K., Ito, T.: Compromising strategy based on estimated maximum utility for automated negotiation agents com-petition (anac-10). In: Modern Approaches in Applied Intelligence, pp. 501–510. Springer, Berlin (2011)
17. Sycara, K.P.: Argumentation: planning other agents'plans. In: Proceedings on International Joint Conference on Artificial Intelligence (IJCAI-89), pp. 517–523 (1989)
18. Sycara-Cyranski, K.: Arguments of persuasion in labor mediation. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-85), pp. 294–296 (1985)
19. Sierra, C., Jennings, N.R., Noriega, P., Parsons, S.: A framework for argumentation-based negotiation. In: ATAL (1997)
20. Wakaki, T., Nitta, K.: Mathematical discussion. Tokyo Denki University Press(Japanese) (2017)
21. Yokoo, M., Durfee, E.H., Ishida, T., Kuwabara, K.: The distributed constraint satisfaction problem: formalization and algorithms. IEEE Trans. Knowl. Data Eng. **10**(5), 673–685 (1998)
22. Fioretto, F., Pontelli, E., Yeoh, W.: Distributed constraint optimization problems and applications: A survey. J. Artif. Intell. Res. **61**, 623–698 (2018)
23. Rubinstein, A.: Perfect equilibrium in a bargaining model **50**(1), 97–109 (1982)

Software Developer Performance Measurement Based on Code Smells in Distributed Version Control System

Natach Jongprasit and Twittie Senivongse

Abstract Effectively staffing a software development team for a software project is important to the development of software in the project. Performance of software developers can be measured by the quality of the produced software, and the number of code smells is one factor that indicates software quality. However, modern software development uses distributed version control systems in which different software developers collaborate to develop software. The number of code smells in the software is hence the result of the aggregate performance of the whole team. This makes it difficult to measure the performance of individual developers. This paper proposes a method and a supporting tool for measuring the performance of individual software developers in a Git project based on code smells. Bayesian average rating is adopted to rate the performance of each developer in a project by taking into account his/her level of contribution to the project, i.e. the number of source code commits, as well as his/her effort to produce clean code. An experiment on C# projects shows that there is a strong positive correlation between ranking of developer performance in a project by the proposed method and that by human evaluators.

Keywords Software developer performance · Code smell · Bayesian average rating · Git · Distributed version control system

1 Introduction

Effectively staffing a software development team for a software project is an important factor for project success. While there are effective methods for measuring team performance, leading experts agree that it is hard to find objective measures

N. Jongprasit · T. Senivongse (✉)
Department of Computer Engineering, Faculty of Engineering,
Chulalongkorn University, Bangkok, Thailand
e-mail: twittie.s@chula.ac.th

N. Jongprasit
e-mail: natach.j@student.chula.ac.th

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_6

for individual performance [1]. Quantitative measurement of individual software developer performance has been controversial since some developers argue that the measurement is likely not to capture everything that is involved and could have negative impact on developers motivation and productivity [2]. From a different point of view, measuring the performance of a software developer can give more insight into the developer quality and allow the developer and project manager to see both the strong points and weak points to be improved. A project manager may assign developers to a software project by conducting a developer ability test or having a team discussion about who have the necessary skills and fit the project. Software professionals have discussed several productivity and code quality metrics that may be used to determine individual developer performance. Some of them are qualitative while some are quantitative. Those metrics include speed and amount of work done [1, 2], test coverage [1, 2], number of reported bugs [1, 2], number of problems found in code reviews [1], codebase understanding [1], adherence to coding standards [1], technical debt in the code [2], responsibility for the good and bad of the code [1], and constant learning and skill improvement [1].

The number of bad smells in the code can be considered as one of the code quality aspects of the developer performance. Code smells or bad smells refer to symptoms of poor design and implementation choices in the code [3]. The motivation of this work is that less number of code smells contributes to better code quality and a software developer who produces code with less number of smells is likely to have good technical performance. However, modern software development utilizes distributed version control systems (DVCS) in which several software developers collaborate in a software project. A popular DVCS system, i.e. Git [4], records changes to files as different versions over time and allows developers to check out a specific version and clone the full history of the file repository onto their local machines. In this collaboration setting, all developers who previously have committed changes to a source code file may have their share of the code smells found in a particular version of the file. As the number of code smells in any version of the file is a result of the aggregate performance of the whole team, how to measure the performance of individual developers in this collaboration setting becomes a challenge.

This paper proposes a method and a supporting tool for measuring the performance of individual software developers in a Git project based on code smells. Bayesian average rating [5] which is used for rating items, e.g. product items on e-commerce web sites, is adopted to calculate the performance of each developer in a project. As Bayesian average considers both the number of votes an item receives and the rating score, it is more appropriate than simple average rating. Likewise, the calculation of individual developer performance takes into account the developers level of contribution to the project, i.e. the number of source code commits, as well as the developers effort to produce clean code in each commit. In an experiment, code smells in C# projects on GitHub [6] were detected by a smell detection tool called Designite [7], and developer performance was calculated. The correlation between

Table 1 Implementation smells supported by Designite [7]

Implementation smell	Description
Complex conditional	A complex conditional statement
Complex method	A method with high cyclomatic complexity
Duplicate code	A code clone within a method
Empty catch block	A catch block of an exception is empty
Long identifier	An identifier with excessive length
Long method	A method is excessively long
Long parameter list	A method has long parameter list
Long statement	An excessively long statement
Magic number	An unexplained number is used in an expression
Missing default	A switch statement does not contain a default case
Virtual method call from constructor	A constructor calls a virtual method

ranking of developer performance in a project by the proposed method and that by human evaluators was tested.

The rest of the paper is organized as follows. Section 2 discusses background and related work. Section 3 proposes the method and a supporting tool to calculate performance of individual developers. An experiment to evaluate the method is presented in Sect. 4, followed by the conclusion in Sect. 5.

2 Background and Related Work

2.1 Code Smells

Code smells are symptoms of poor design and implementation choices in the code which may indicate deeper problems and affect program maintainability [3]. Code smells are not bugs as code still functions but they may lead to bugs and failure in the future. Here, the focus is on two categories of code smells: (1) implementation smells [3] which are code structures that indicate potential problems in the implementation, and (2) design smells [8] which are code structures that violate fundamental design principles. In the experiment in this paper, the smell detection tool called Designite [7] was used to detect code smells in source code written in C#. Designite can support detection of 11 implementation smells and 19 design smells as listed in Tables 1 and 2.

Table 2 Design smells supported by Designite [7]

Design smell	Description
Broken hierarchy	A supertype and its subtype conceptually do not share an IS-A relationship
Broken modularization	Data and/or methods that ideally should have been localized into a single abstraction are separated and spread across multiple abstractions
Cyclically-dependent modularization	Two or more abstractions depend on each other directly and indirectly
Cyclic hierarchy	A supertype in a hierarchy depends on any of its subtype
Deep hierarchy	An inheritance hierarchy is excessively deep
Deficient encapsulation	The declared accessibility of one or more members of an abstraction is more permissive than actually required
Duplicate abstraction	Two or more abstractions have identical names or identical implementation
Hub-like modularization	An abstraction has high incoming and outgoing dependencies
Imperative abstraction	An operation is turned into a class
Insufficient modularization	An abstraction exists that has not been completely decomposed, and a further decomposition could reduce its size or implementation complexity
Missing hierarchy	Conditional logic to explicitly manage variation in behavior
Multifaceted abstraction	An abstraction has more than one responsibility assigned to it
Multipath hierarchy	A subtype inherits both directly as well as indirectly from a supertype
Rebellious hierarchy	A subtype rejects the methods provided by its supertype(s)
Unexploited encapsulation	Client code uses explicit type checks
Unfactored hierarchy	There is unnecessary duplication among types in a hierarchy
Unnecessary abstraction	An abstraction that is actually not needed
Unutilized abstraction	An abstraction is left unused
Wide hierarchy	An inheritance hierarchy is too wide

2.2 Distributed Version Control System

To support collaboration between software developers in a software project, a distributed version control system such as Git [4] has a server that contains all versioned

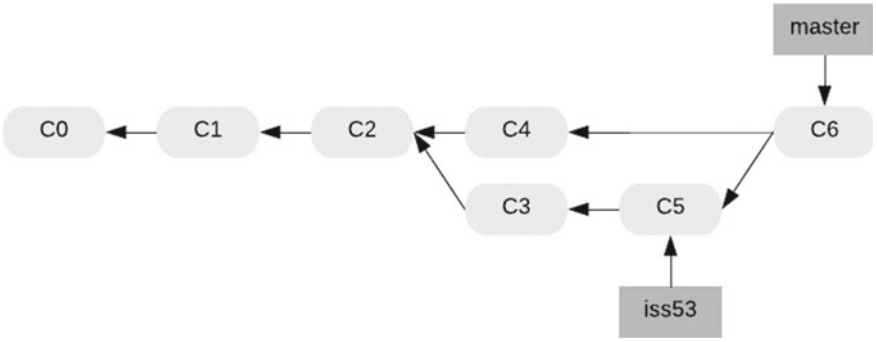


Fig. 1 Branching off and merging (or pulling) into the master [4]

files of the project. Developers can fully mirror or clone the whole project locally and then check out the snapshot of the files they want to modify. Since developers have access to the full history of the project, communication with the server is reduced to only when the modified files need to be merged back to the server. If the server dies, the mirrored copy can be used to restore the server. In this paper, an experiment was conducted on C# projects on GitHub [6]—a popular web-based code hosting platform built on Git software. GitHub provides an easy-to-use interface for developers to create a repository, or Git project, to contain folders and files (e.g. source code, images, videos, documents, and data sets), together with each files history, i.e. a list of file commits in time [9]. By default, a repository has one master (or production) branch of development. Different developers can create other branches off the master branch to make edits to the project before merging them into the master. Figure 1 shows a commit history in the master and in a branch called iss53 that later is merged (or pulled from a remote repository to merge) into the master at commit C6. In addition, GitHub allows for comparing differences between versions of the code snapshots. In this paper, code smell detection is performed on code snapshots in the master and branches of the project.

2.3 Bayesian Average Rating

Bayesian average rating is used commonly by web sites to determine an average score for an item, e.g. product or service, based on rating by web site visitors [5]. The scores are then used to rank the items. Bayesian average is considered more believable than simple average when rating items since it also considers the number of votes each item has received. For example, suppose an item receives 1 point for each positive vote and 0 point for each negative vote. Using simple average, an item that receives 95 positive votes and 5 negative votes out of 100 votes would have a

rating of 0.95, whereas another item that receives 2 positive votes out of 2 votes would have a rating of 1.0 and would be ranked first. However, many votes should count more, or have higher weight, than few votes as they reflect how most people feel about the item. Hence, the relation of item ratings to each other is considered based on the number of votes of each item as well as the score each item receives. With Bayesian average rating, the less votes an item has, the closer its rating should be to the average rating of all items. More new votes to an item pull its rating from the average toward a more believable value. Bayesian average rating of a particular item is calculated by

$$br = \frac{(avg\_num\_votes \times avg\_rating) + (item\_num\_votes \times item\_rating)}{avg\_num\_votes + item\_num\_votes} \tag{1}$$

where br = Bayesian average rating of this item (recalculated when there is a new vote for any item),

avg_num_votes = average number of votes of all items,
avg_rating = average rating score of all items,
$item\_num\_votes$ = number of votes of this item,
$item\_rating$ = average rating score of this item.

In this paper, the calculation of individual developer performance applies Bayesian average rating by taking into account the number of code commits made by each developer as well as the score each developer receives from fixing code smells in each commit.

2.4 Related Work

An empirical study by Sharma et al. [10] investigated the characteristics of code smells in 1,988 C# repositories on GitHub. Designite was used to detect 11 implementation smells and 19 design smells. It was found that (1) magic number and unutilized abstraction were the most frequently occurring smells in C# code, (2) there was a strong positive correlation between the occurrence of implementation smells and design smells, (3) magic number and unutilized abstraction had the highest co-occurrence with other implementaion smells and design smells respectively, and (4) there was a weak correlation between smell density and line of code, and hence it was undecidable whether smell density was associated with the size of the project. Li et al. [11] proposed a method to extract characteristics of software developers based on the history of collaboration in Git projects on GitHub which used MVC architecture. The characteristics included the area of contribution (i.e. model, view, or controller layers), level of contribution based on frequency of changes made

to the files, and level of initiative based on frequency of creation of system environment files and tools, and roles (i.e. support or leader). This work is close to ours in that it also analyzed developer behavior from Git projects, but it did not determine developer performance. Alnaji and Salameh [12] proposed a performance measurement framework for software engineers who used agile methods such as scrum. Measures that were considered included (1) productivity based on story points completed, (2) efficiency based on handling of defects, time to complete user stories, and spilled-over story points onto future sprints, (3) social skill, (4) mentorship and team collaboration, and (5) breadth of knowledge. This framework focused more on team performance, rather than individual performance.

3 Method to Measure Performance of Individual Software Developer

Measuring individual performance of each software developer who collaborates in a Git project is based on the quality (i.e. number of smells) of the code committed by each developer as well as the number of code commits by each developer. For each source code commit made by a developer, the measurement comprises four steps: (1) code smell detection in code snapshot, (2) code smell density calculation, (3) performance score calculation based on previous code snapshot, and (4) overall performance calculation using Bayesian average rating.

3.1 Code Smell Detection in Code Snapshot

Figure 2 depicts a typical collaboration scenario in a Git project. Developers may clone a project, branch off the master branch to modify the files, and make commits. Later the branch can be pulled to merge with the master when the modification is done. Measuring the performance of a developer considers only the commits made by adding or modifying source code. Commits of other files that are not source code are not considered. Other operation such as branching is not considered either as it does not create a new snapshot of the code. Although a pull, or merge, operation creates a new code snapshot, it does not represent a code contribution of the developer who makes a pull or merge, and hence is not considered in the calculation of his/her performance. In Fig. 2, all six code snapshots (i.e. initial code, commits C1-C4, and code at the pull request) are scanned for code smells. Commits C1 and C3 made by a developer D1 will be used to calculate the performance of D1, whereas commits C2 and C4 will be used in the performance calculation of another developer D2. Note that, the initial codebase in this example is not created by the current development

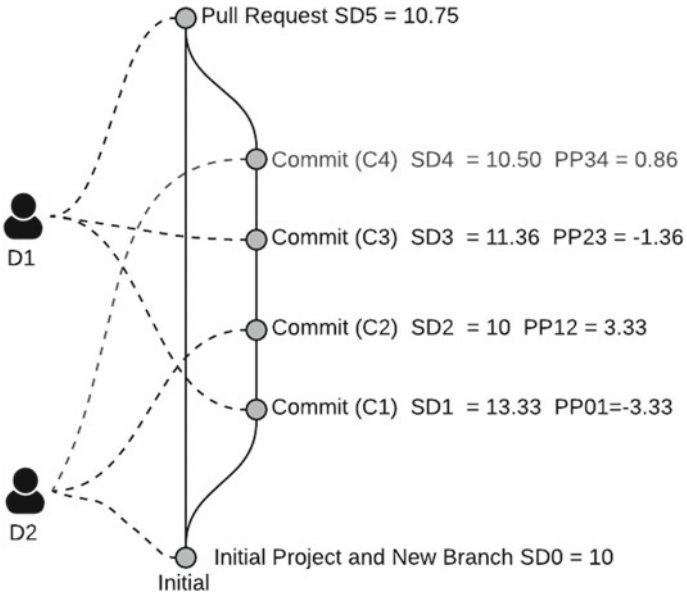


Fig. 2 Example of history of code snapshots for code smell detection

team and is not a contribution of any developer in the project. Also, although D1 pulls the branch, the resulting code snapshot is not considered D1's contribution. Anyway, the initial codebase and the code snapshot created by the pull request are scanned, so that the detected code smells will be used further to calculate developer performance.

3.2 Code Smell Density Calculation

In this paper, the quality of a code snapshot is determined by code smell density [10], i.e. the number of code smells in the code snapshot in relation to the size of the code. Code smell density is calculated by

$$SD_j = \frac{Number\ of\ smells\ in\ code_j}{KLOC_j} \tag{2}$$

where j = a version of code snapshot,

SD_j = smell density of a code snapshot j,
$Number\ of\ smells\ in\ code_j$ = number of code smells in code snapshot j,
$KLOC_j$ = size of code snapshot j (in the unit of a thousand line of code).

Suppose the smell density SD_j of all six versions of the code snapshots are as shown in Fig. 2. For example, if the code snapshot at commit C4 has 25 code smells

and the size is 2.38 $KLOC$, the smell density SD_4 is $25/2.38 = 10.5$ smells per $KLOC$.

3.3 Performance Score Calculation Based on Previous Code Snapshot

In a collaboration setting, all developers who previously have committed changes to the source code may have their share of the smells found in a particular code snapshot. Therefore, the performance of the developer who commits that code snapshot is based on whether he/she has improved the quality of the code, compared with the previous version of the code snapshot. That is, the developer would be considered as having good performance if he/she can reduce the smell density by fixing some smells in the previous version of the code and not adding many new smells to the added and modified code in the version that he/she has committed. The performance score of a developer based on previous code snapshot is therefore calculated by

$$PP_{ij} = SD_i - SD_j \tag{3}$$

where j = a version of code snapshot committed by this developer,

i = a previous version of code snapshot j,
PP_{ij} = performance score of this developer based on quality improvement made to the code snapshot j in comparison with the previous version i,
SD_j = smell density of a code snapshot j,
SD_i = smell density of a previous code snapshot i.

In Fig. 2, the score PP_{34} of the developer D2 who makes a commit C4 is 0.86 (i.e. $SD_3 - SD_4 = 11.36 - 10.50$). The value is positive, meaning that D2's performance is good in that he/she has cleaned up some smells when modifying the commit C3 before committing C4. On the other hand, the score PP_{23} of the developer D1 who makes a commits C3 is -1.36 (i.e. $SD_2 - SD_3 = 10 - 11.36$). The value is negative, meaning that D1's performance is not so good since the modification of the code at commit C2 by D1 has introduced more smells to the code at commit C3. In the case that PP_{ij} is 0, the developer who commits the code snapshot j does not make any change to the quality of the code snapshot i.

3.4 Overall Performance Calculation Using Bayesian Average Rating

Different developers have different levels of contribution to the project. Developer habits and skills in producing code of good or poor quality (i.e. clean code or smelly

code) could be observed better when the developer contributes greatly to the project and make a long history of commits. The calculation of the overall performance of a developer throughout the project adopts Bayesian average rating so that the developers level of contribution, i.e. the number of source code commits, and the developers effort to produce clean code in each commit are considered. The overall performance of a developer in a project is computed by

$$PB_d = \frac{(avg\_num\_comm \times avg\_perf) + (dev\_num\_comm \times dev\_perf)}{avg\_num\_comm + dev\_num\_comm} \tag{4}$$

where PB_d = overall performance of a developer d in a project (recalculated when there is a new commit by any developer),

avg_num_comm = average number of code snapshots committed by all developers in this project

$$= \frac{No.\ of\ code\ snapshots\ committed\ in\ project}{No.\ of\ developers\ in\ project}, \tag{5}$$

avg_perf = average performance score of all developers in this project

$$= \frac{\sum PP_{ij}\ for\ code\ snapshots\ committed\ in\ project}{No.\ of\ code\ snapshots\ committed\ in\ project}, \tag{6}$$

dev_num_comm = number of code snapshots committed in this project by the developer d,

dev_perf = average performance score of the developer d in this project

$$= \frac{\sum PP_{ij}\ for\ code\ snapshots\ committed\ in\ project\ by\ developer\ d}{No.\ of\ code\ snapshots\ committed\ in\ project\ by\ developer\ d}. \tag{7}$$

In Fig. 2, there are two developers, D1 and D2, in the project. There are four code snapshots, C1–C4, that are modified and committed. Both developers make two commits. PP_{ij} scores are as shown in the figure. Thus, the overall performance score of D2 in this project (upto the latest commit C4) is computed as follows:

$avg\_num\_comm = 4/2 = 2,$
$avg\_perf = ((-3.33) + 3.33 + (-1.36) + 0.86)/4 = -0.125,$
$dev\_num\_comm = 2,$
$dev\_perf = (3.33 + 0.86)/2 = 2.095.$
Then $PB_{D2} = ((2 * (-0.125)) + (2 * 2.095))/(2 + 2) = 0.985.$

The overall performance score of D1 in this project (upto the latest commit C4) can be computed in the same manner and the PB_{D1} score is -1.235. Therefore, D2 performs better than D1 in this project as D2 is likely to improve the code quality by committing cleaner code.

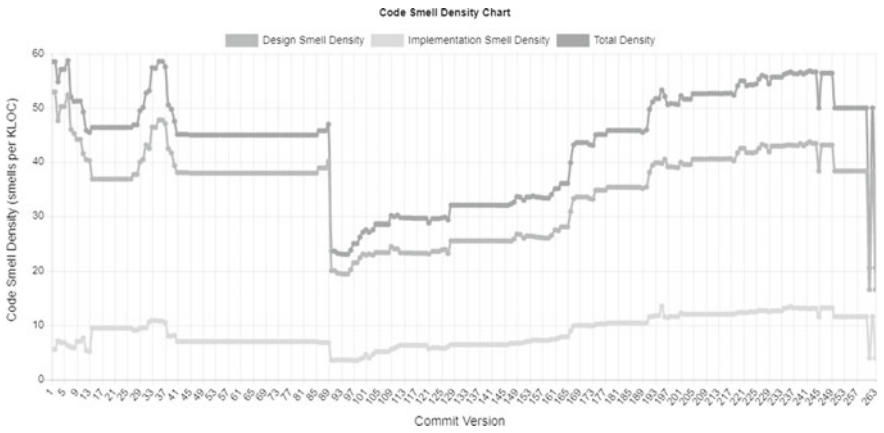


Fig. 3 Code smell density in a C# project on GitHub

3.5 Supporting Tool

A web-based tool, written in C#, has been developed to support the proposed method. A user (e.g. project manager or senior team member) who wants to evaluate individual developer performance can add a project to the tool by specifying the name of a C# Git project and the project owner. The tool then retrieves project information from GitHub including project history, all commits, and committers. For each version of the code snapshots, the tool calls Designite to detect code smells and then calculates the overall performance of each developer in the project. The user can also refresh project information to retrieve new commits and have the performance scores recalculated. The tool can report smell density of each code version by smell categories as in Fig. 3, and the number of commits made by each developer as in Fig. 4. Particularly, the tool can compare the overall performance of each developer in the project. Figure 5 shows the overall performance of each developer when the score calculation is based on the ability to reduce design smells, implementation smells, and any types of smells. This implies that, even though the previous code version that a developer is modifying is poor and contains many smells, if the developer can refactor some of the existing smells or the code that is added is clean, the smell density would be reduced in the next commit. It reflects this developer's ability to improve the code and hence the performance is good in relation to the previous code snapshot. Note that the individual performance scores should be used for comparison within the same project. Different projects differ in environment, complexity, size, and team. Care must be taken if the method will be used to determine the performance of a developer across different projects or compare the performance of developers in different projects since those factors can affect the performance scores.

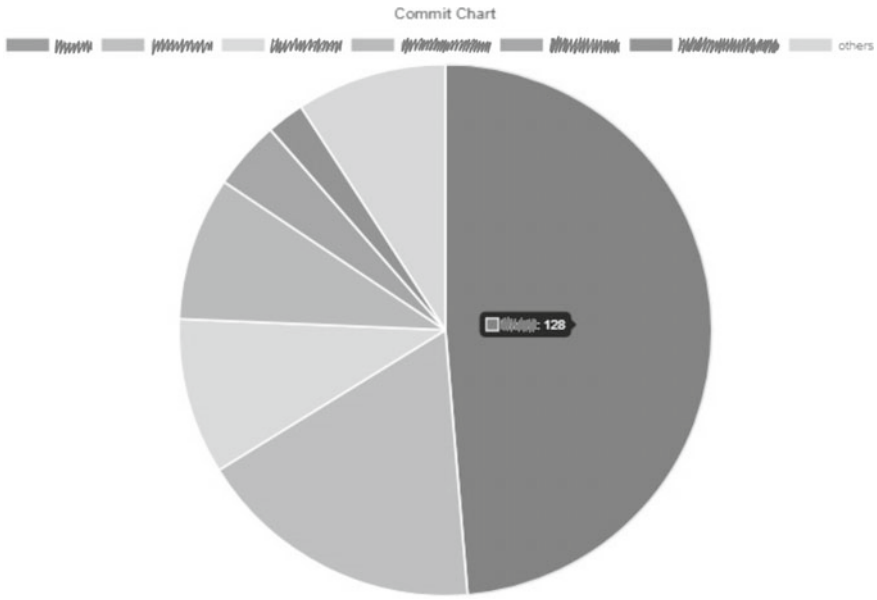


Fig. 4 Number of commits made by each developer in a C# project on GitHub (developer names are concealed)

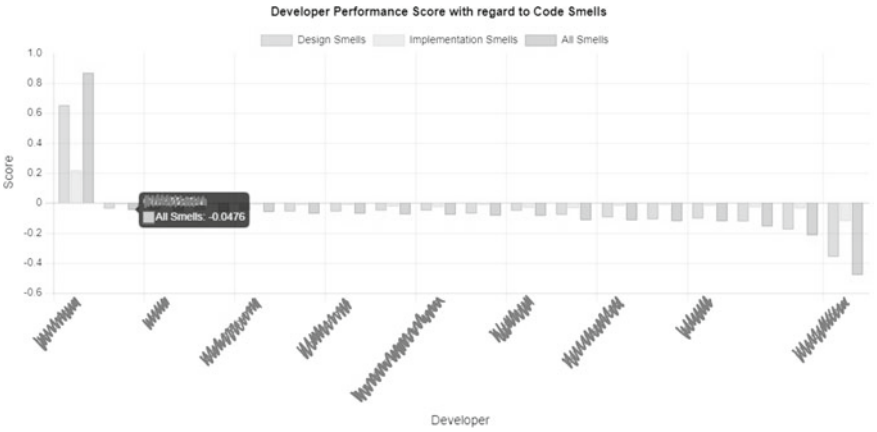


Fig. 5 Developer performance scores with regard to code smells in a C# project on GitHub (developer names are concealed)

4 Evaluation

To evaluate if the measurements from the proposed method can represent the performance of developers in the project, an experiment was conducted to test if there is a correlation between the ranking of developer performance in a project as measured by the tool and the ranking by evaluators.

4.1 Experiment

Four evaluators were asked to go through the code history of two C# projects on GitHub and study the added or modified code in the commits made by each developer. Three evaluators have 6-year and one has 2-year experience in the senior programmer position. The detail of the two projects are in Table 3. After looking at the code, each evaluator rated each developer by the criteria in Table 4 with the score ranging from 1 (need improvement) to 5 (very good). After that, the total score of each developer in each project, given by each evaluator, was computed. Then Spearman's

Table 3 Detail of experimental C# projects

Project/detail	Project 1	Project 2
No. of commits	292	82
Line of code	5482	3396
No. of developers	18	7

Table 4 Criteria to evaluate developer performance

Does the code added or modified by the developer have these characteristics?
1. Code is well-structured, e.g. separated into layers, having appropriate calls between classes
2. Identifier names, e.g. class, method, and variable names, are meaningful and easy to understand
3. Different functions are clearly assigned to different methods
4. Classes and methods are not too large and are easy to understand
5. Methods having related functions belong to the same class
6. Code is properly organized for reuse
7. Magic numbers are properly handled, e.g. using constant declaration instead
8. Comments are given when necessary, e.g. to explain complex algorithm or reason for implementation
9. Old code is properly handled, e.g. removed if not used
10. Code that is at risk of having problems is properly managed, e.g. potential bugs are fixed and try-catch blocks are put in appropriate places

Table 5 Spearman's rank correlation coefficients

Tool—Evaluator	r of project 1	r of project 2
Tool—Evaluator 1	0.9322	0.8321
Tool—Evaluator 2	0.9228	0.9405
Tool—Evaluator 3	0.8915	0.8321
Tool—Evaluator 4	0.9749	0.9303
Average	0.9304	0.8838

rank correlation coefficients (r) were calculated between the ranking of the scores by the evaluators and the ranking of the scores by the supporting tool.

4.2 Result and Discussion

The Spearman's rank correlation coefficients are listed in Table 5. The average coefficients are 0.9304 and 0.8838 in project 1 and project 2 respectively, indicating a strong positive association between the ranking of the scores by the tool and that by all evaluators in both projects.

A threat to validity of the experiment could be that, for developers who have very few commits in the project and do not add or change a lot to the code, there is not much information for the evaluators to judge them by the criteria. In that case, they would be rated with a moderate score (i.e. 3). This happens to align with how Bayesian average rating works, i.e. if the developer does not contribute much, the score would be pushed toward the average performance score of the project. This could affect the experiment and make the correlation coefficients high.

5 Conclusion

This paper proposes a method and a set of metrics, based on Bayesian average rating, to compute individual performance of developers in a software project in which the collaboration is via a distributed version control system like Git. The method considers the skills of the developers in improving code quality by reducing code smell density in the code they commit in the project. The experiment shows that the method has a strong positive correlation with the performance evaluation by human evaluators, and hence the measurements could represent the performance of each developer. Although the proposed method enables direct objective measurement of individual performance, the method should be used to complement other subjective or objective methods since individual performance evaluation should consider different aspects of a developer all round.

It is expected that the proposed method can be applied to projects that use other distributed version control systems or written in other programming languages. As the key idea of the method is to determine developer performance from change of smell density between different commits in a collaborative development setting, the method can also be applied in a similar manner to other version control environment including centralized version control systems. The method can be enhanced to consider other aspects of code quality, apart from code smells, such as maintainability and understandability which can be determined from other properties of the code.

References

1. Hodges, N.: Can developer productivity be measured? https://dev.to/nickhodges/can-developer-productivity-be-measured-1npo. Accessed 19 Aug 2018
2. York, B.: The best developer performance metrics. https://medium.com/@yupyork/the-best-developer-performance-metrics-6295ea8d87c0. Accessed 16 Aug 2015
3. Fowler, M., Beck, K., Brant, J., Opdyke, W., Roberts, D.: Refactoring: Improving the Design of Existing Code. Addison-Wesley, Boston (1999)
4. Chacon, S., Straub, B.: Pro Git, 2nd edn. Apress, Berkeley. https://git-scm.com/book/en/v2 (2014)
5. Weichselbaum, M.: Bayesian rating how to implement a weighted rating system. http://thebroth.com/blog/118/bayesian-rating.html. Accessed 30 March 2006
6. GitHub. https://github.com
7. Sharma, T.: Designite: a customizable tool for smell mining in C# repositories. In: Seminar Series on Advanced Techniques and Tools for Software Evolution (SATToSE 2017), 5 p.
8. Suryanarayana, G., Samarthyam, G., Sharma, T.: Refactoring for Software Design Smells: Managing Technical Debt. Morgan Kaufmann, San Francisco (2015)
9. GitHub Guide, Git Handbook. https://guides.github.com/introduction/git-handbook
10. Sharma, T., Fragkoulis, M., Spinellis, D.: House of cards: code smells in open-source C# repositories. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 424–429 (2017)
11. Li, S., Tsukiji, H., Takano, K.: Analysis of software developer activity on a distributed version control system. In: 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 701–707 (2016)
12. Alnaji, L., Salameh, H.: Performance-measurement framework to evaluate software engineers for agile software-development methodology. Eur. J. Bus. Manag. **7**(2), 183–190 (2015)

Decision Support System for Choosing an Elective Course Using Naive Bayes Classifier

Abiyoga, Arya Wicaksana and Ni Made Satvika Iswari

Abstract Department of Informatics in Universitas Multimedia Nusantara offers five elective courses to the students. Choosing an elective course that is most suitable for the student interest and academic skills is dilemmatic. Thus, a decision support system is proposed to assist the students in choosing an elective course based on not only their interest but also academic skills. The system uses Naive Bayes Classifier and Laplace smoothing for the classification process. The data used for this research is collected from 120 students. Learning from past students records, the system could predict the outcome of the student upon choosing an elective course. The evaluation of the system shows that the accuracy of the system is 0.30 and 0.33, recall is 0.318 and 0.378, precision is 0.215 and 0.407, and the F-score are 0.257 and 0.390. A test for two classes and three classes classification using 60 generated data shows improvement in the performance with the accuracy of 0.83 and 0.72 and the F-score of 0.843 and 0.728.

1 Introduction

Department of Informatics in Universitas Multimedia Nusantara (UMN) offers five elective courses for the students. The elective course is to be taken at the 5th, 6th, and 7th semester. These five elective courses available for the students to choose are:

- Game Design and Development
- Cyber Security
- Database Administration

Abiyoga · A. Wicaksana (✉) · N. M. S. Iswari
Department of Informatics, Universitas Multimedia Nusantara, Tangerang 15810, Indonesia
e-mail: arya.wicaksana@umn.ac.id

Abiyoga
e-mail: abiyoga@student.umn.ac.id

N. M. S. Iswari
e-mail: satvika@umn.ac.id

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_7

- Applied Computer Networking
- System Applications Products (SAP)

These elective courses are intended to broaden students knowledge and skills in a particular field. Choosing one from five available elective courses that are most suitable with student interest and academic skills is dilemmatic. Another consideration to take is the prediction of the outcome that the students could achieve for each elective course before choosing one.

A decision support system (DSS) is a system that uses flexible, interactive, and adaptable Computer Based Information System (CBIS), to support a solution for unstructured specific management problem [1]. Decision support systems are usually built to support solutions to a problem or to an opportunity. A DSS allows users sift through and analyze great amounts of data, and compile information that could be used to solve problems and generate better decisions. Thus, in this case, students academic records are used to help the system recommends the best elective course for them by comparing the academic scores with the previous batch of students. Hence, the prediction is made by learning the outcomes of past students scores for each elective course.

The Naive Bayes Classifier (NBC) is used in this work for the DSS to make a recommendation. NBC is a data mining tool. Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning to extract and identify useful and related information from various large databases. Many functions could be applied from data mining, including estimation, prediction, clustering, classification, and association. Naive Bayes Classifier is a simple probabilistic classification that calculates a set of probabilities by summing frequencies and combinations of values from a given dataset. Naive Bayes Classifier is based on the simplifying assumption that attribute values are conditionally independent. The advantage of using Naive Bayes Classifier is that this method only requires a small amount of training data (training data) to determine the estimation of parameters required in the classification process.

Based on the research in [2], several algorithms such as Neural Networks (NN), Naive Bayes (NB), and Decision Tree (DT) are compared for automatic analysis and classification of attribute data from web pages. The result is that Naive Bayes is the best algorithm compared to NN and DT with a value of F-score of more than 97%. It is shown in the research, that Naive Bayes Classifier is easier to use because it has a long calculation flow whereas the Decision Tree Algorithm (C4.5) takes longer calculation when the data is changed or added.

In this work, the Naive Bayes Classifier is used for the DSS to classify and predict the most suitable elective course. The recommended elective course is generated by the system based on student interest and academic skills. The data used for training and testing the classifier are obtained from past 120 student academic records. These

past records are useful for the system to predict the student outcome for each elective course. The proposed system is developed as a web-based application using PHP and connected to a MySQL database.

2 Methods

2.1 Decision Support System

The decision support system (DSS) could be described as a system that generates information aimed at a particular problem that has to be solved by the manager and support in decision-making [1]. The DSS consists of three interacting components [3]:

1. The knowledge system (problem domain knowledge repositories that exist in a decision support system or as data or a procedure)
2. The problem processing system (consist of one or more capabilities of common problem manipulation needed for decision making)
3. The language system (a mechanism to provide communication between users and other decision support system components)

The three main components above could be realized into [1]:

1. Database: The component which is useful as a data provider for the system. The data is kept and organized in a system called database management system.
2. Model base: The imitation of the real world. The obstacles that often come out in designing the model are when the designed model is not able to reflect all the variable of the real world so that the decisions taken do not suit the needs.
3. User interface management system: The facility that is able to integrate the installed system with the users interactively.

2.2 Naive Bayes Classifier

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. Naive Bayes classifier (NBC) is a simple probabilistic classification that calculates a group of probabilities by adding up the frequency and combination of values from the given dataset. The algorithm assumes all the attributes given by value in class variables are independent, in other words, not interdependent [4, 5], therefore the calculation can be done as the following formula:

$$P(c|x) = \frac{P(x|c) \times P(c)}{P(x)} \tag{1}$$

where:

$P(c|x)$ = the posterior probability of class (target) given predictor (attribute)

$P(c)$ = the prior probability of class

$P(x|c)$ = the likelihood which is the probability of predictor given class

$P(x)$ = the prior probability of predictor

P(x) could be ignored when P(x) is identical for all classes. Thus, the Naive Bayes Classifier could be defined by discriminant function as the following [5].

$$NB(x) = \prod_{j=1}^{n} P(X_j = x_j | C = i) P(C = i) \tag{2}$$

where:

$P(X_j = x_j|C = i)$ = the probability of data X given that the hypothesis C was true

$P(C = i)$ = the probability of hypothesis C being true

There are cases where some of the parameters would not be in training sets and the value is 0. The multiplication of 0 with others would be 0. This could be avoided by using Laplace smoothing or add-one smoothing. The use of Laplace smoothing with the NBC here is to add 1 to every count so it would not be 0.

2.3 Confusion Matrix

The confusion matrix is a table that records the results of classification work. Figure 1 shows the example of a confusion matrix that classifies binary problems (two classes) for two classes: 0 and 1 [6].

True positive (TP) is the number of positive records classified as positives, false positives (FP) is the number of positive records classified as negatives, false negatives (FN) is the number of negative records classified as positives, true negatives (TN) is the number of negative records classified as negatives (Fig. 2).

The following formula are used to find the total number of FN, FP, TN for each class i respectively. The total true positive in the system will be obtained through Formula 6 [7].

Fig. 1 Confusion matrix for 2 classes

f_{ij}		Predicted	
		Class 1	Class 2
Actual	Class 1	TP	FN
	Class 2	FP	TN

Fig. 2 Confusion matrix for multi-classes [7]

		Predicted Number			
		Class 1	Class 2	…	Class *n*
Actual Number	Class 1	x_{11}	x_{12}	…	x_{1n}
	Class 2	x_{21}	x_{22}	…	x_{2n}
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	Class *n*	x_{n1}	x_{n2}	…	x_{nn}

$$TFN_i = \sum_{j=1, j\neq 1}^{n} x_{ij} \tag{3}$$

$$TFP_i = \sum_{j=1, j\neq 1}^{n} x_{ji} \tag{4}$$

$$TTN_i = \sum_{j=1, j\neq 1}^{n} \sum_{k=1, k\neq 1} x_{jk} \tag{5}$$

$$TTP_{all} = \sum_{j=1}^{n} x_{jj} \tag{6}$$

The quantities that could be used as classification performance metrics other than accuracy are precision and recall. Both of these quantities provide more relevant performance values [6]. Recall (true positive rate) measures the original positive proportion that is correctly recognized (predicted). Precision measures how often the proportion of true positive predictions are right. The following formula are used to find recall, precision, and overall accuracy [8].

$$Precision_i = \frac{TTP_{all}}{TTP_{all} + TFP_i} \tag{7}$$

$$Recall_i = \frac{TTP_{all}}{TTP_{all} + TFN_i} \tag{8}$$

$$Overall\ Accuracy = \frac{TTP_{all}}{N} \tag{9}$$

The F-score is the harmonic value or the average value between the values of precision and recall. It could be calculated using the following formula [6, 9].

$$F\text{-}score = \frac{2 \times (precision \times recall)}{precision + recall} \tag{10}$$

The value of the performance metrics for classification (accuracy, precision, recall and F-score) could be categorized into five groups [10]:

- 0.90–1.00 = very good classification
- 0.80–0.90 = good classification
- 0.70–0.80 = sufficient classification
- 0.60–0.70 = bad classification
- 0.50–0.60 = wrong classification.

3 Results

3.1 Unirecommend

The DSS is named Unirecommend and Fig. 3 shows the login dialog of the system. The students are registered into the system by the admin of the department. The system will direct the user to the home page as displayed in Fig. 4 upon a successful login. The user interface of the recommendation process is shown in Fig. 5. The process contains text-box for the students to fill in the final scores of each courses fro the 1st to 4th semester. Students are then able to see and save the result. Figure 6 is the user interface to display the recommendation result.

Figure 4 shows the design of the home page. The home page contains sidebar, container and header. The Sidebar consists of two menus: home and logout. The container consists of three buttons: the recommend button (to show a modal dia-

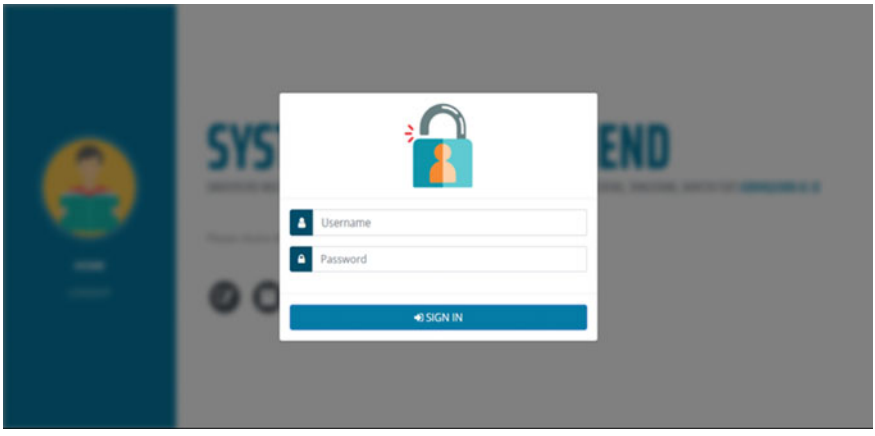


Fig. 3 User interface of recommendation process

Fig. 4 User interface of recommendation process

log recommend), the my recommendation button (to show the capital my recommendation dialog), and the recommendation score button (to show a capital score recommendation dialog). The header contains system information and title.
The translation for each of the subjects in Fig. 5 are:

- Introduction to Multimedia Technology
- Programming Logic
- Discrete Mathematics
- Concept of Programming Languages
- Computers and Society
- Multimedia Concept
- Algorithms and Programming
- Database System
- Operating System
- Data Structure
- Language Theory and Automata
- Human and Computer Interaction
- Computer Graphics and Animation
- System Design
- Object Oriented Programming
- Capita Selecta

In Fig. 5 students could left a text-box empty if a particular subject has not been taken yet. Based on the simulation, the recommended elective course is System Applications Products (SAP). The predicted score of the particular student upon taking the SAP as the elective course could be seen in Figs. 7 and 8. Due to the lack of data distribution for all of the elective courses, not all of the scores could be predicted in the simulation. Another representation of the system confidence level is given in Fig. 8.

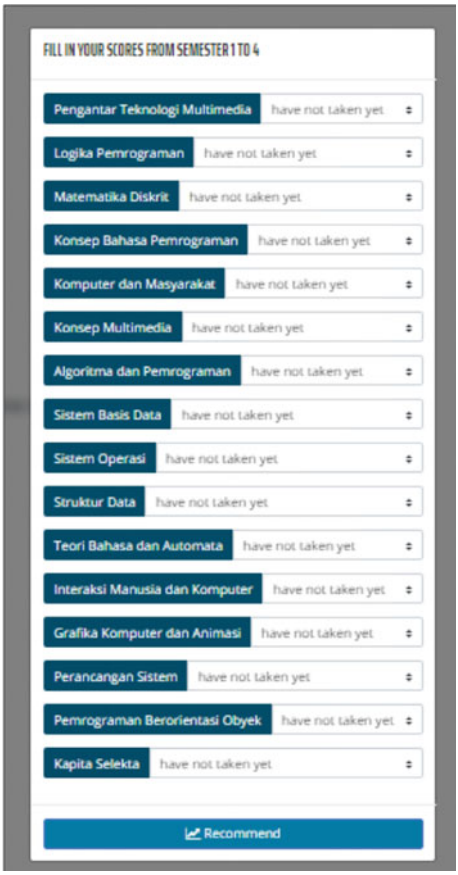


Fig. 5 User interface of recommendation process

3.2 Testing

Testing is conducted to verify the implementation of the Naive Bayes Classifier on the Unirecommend. The first step is to calculate the number of frequency of occurrence of each class: Game Design and Development, Cyber Security, Database Administration, Applied Computer Networking, and System Applications Products (SAP) in the data table and calculate the total number of classes in all categories (number of rows of data) in the data table.

Table 2 shows the number of occurrence frequencies in each class in the training data. The number of times the Database Administration class appears is 30, the number of times the Applied Computer Networking class appears is 32, the number of times the Game Design and Development appears is 13, the number of times the

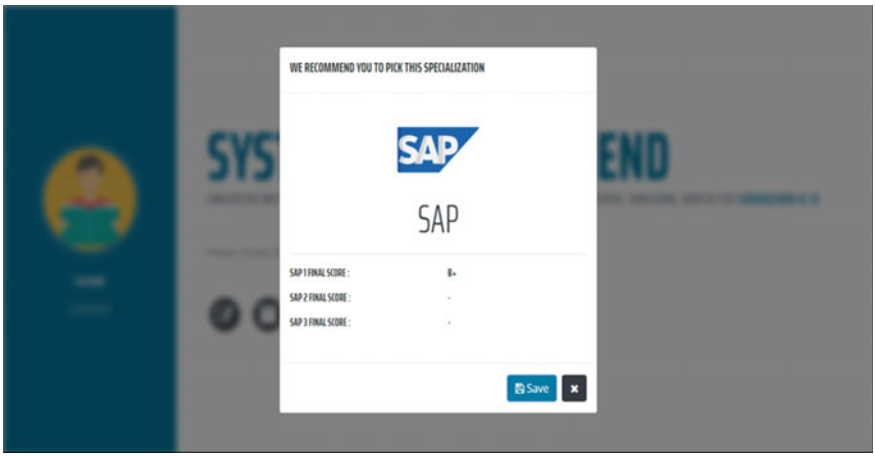


Fig. 6 User interface of recommendation result (1)

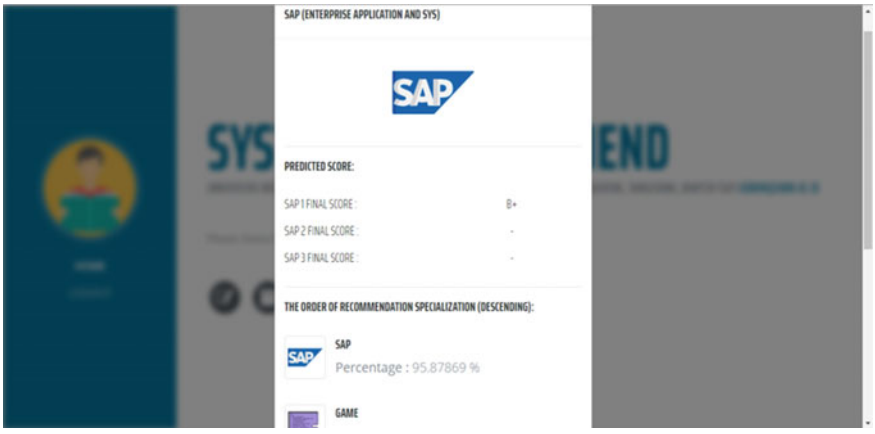


Fig. 7 User interface of recommendation result (2)

Cyber Security class appears is 23 and the number of SAP classes is 22. The total number of data is 120. The parameters used for this test are given in Table 1.

The second step is to calculate the number of frequency of occurrence of each attribute for each class. Table 3 shows the number of frequencies of occurrence of each attribute for each class.

The third step is to calculate the probability value of each attribute to each of the classes. After calculating the probability value of each attribute to each of the classes. The final step is calculating the probability value of each class of all data (prior). The final result of the classification process obtained by manual calculations is the Database Administration with the probability of 7.41959E−14. This is also the same result given by the Unirecommend system as presented in Fig. 9.

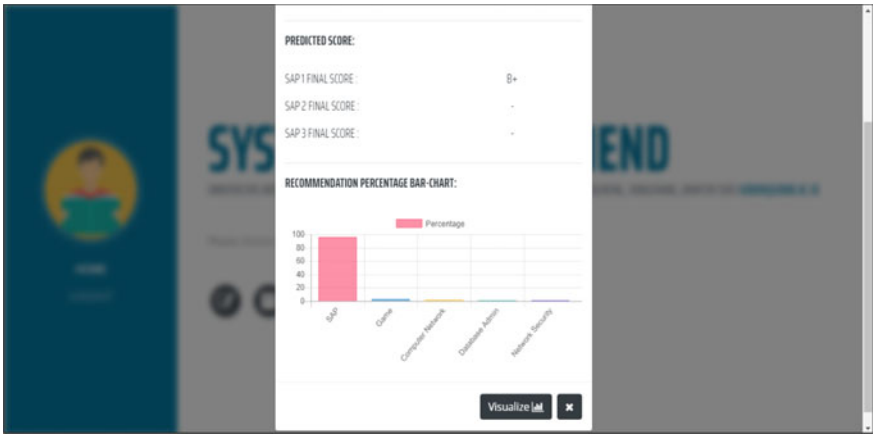


Fig. 8 User interface of recommendation result (3)

Table 1 Test parameters

No	Mandatory courses	Final score
1	Introduction to multimedia technology	A
2	Programming logic	B+
3	Discrete mathematics	B+
4	Concept of programming languages	A
5	Computers and society	A
6	Multimedia concept	A−
7	Algorithms and programming	B−
8	Database system	A−
9	Operating system	B−
10	Data structure	B+
11	Language theory and automata	B+
12	Human and computer interaction	A−
13	Computer graphics and animation	B−
14	System design	A−
15	Object oriented programming	B−
16	Capita selecta	A−

Table 2 Number of frequency of each class

No	Elective courses	Frequency
1	Game design and development	13
2	Cyber security	23
3	Database administration	30
4	Applied computer networking	32
5	System applications products	22

Table 3 Frequency of occurrence of each attribute to each class

No	Attribute	Game	Cyber	Database	Networking	SAP
1	Introduction to multimedia technology	7	15	18	12	2
2	Programming logic	2	5	7	6	5
3	Discrete mathematics	0	2	3	2	4
4	Concept of programming languages	3	8	8	6	0
5	Computers and society	7	8	12	14	5
6	Multimedia concept	4	5	9	11	7
7	Algorithms and programming	1	0	1	8	4
8	Database system	1	5	5	4	1
9	Operating system	2	4	7	8	4
10	Data structure	4	11	13	8	9
11	Language theory and automata	1	4	3	2	3
12	Human and computer interaction	4	11	5	7	2
13	Computer graphics and animation	2	2	6	6	3
14	System design	3	4	7	4	5
15	Object oriented programming	2	1	4	4	1
16	Capita selecta	1	1	1	1	0

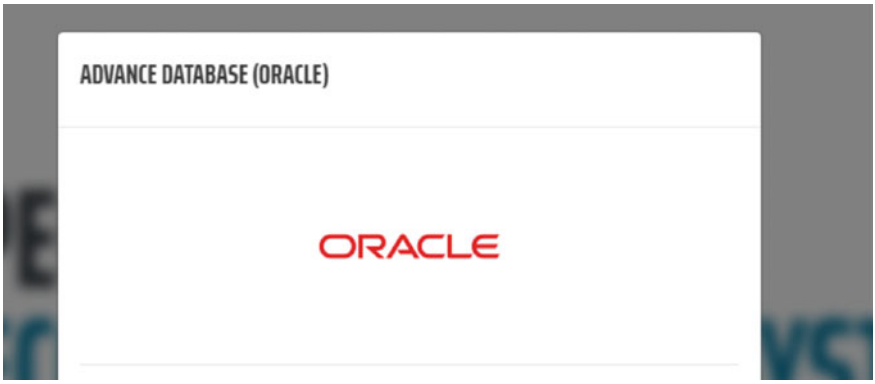


Fig. 9 Testing result

4 Analysis

The total of data used in this work is 120, the proportion of the training data and testing data are divided into 75 and 25%. The evaluation is done using two test-cases. The first test-case uses data number 91–120 for testing. Table 4 shows the accuracy, recall, precision, and F-score of the first test-case.

The second test-case is where the testing data has the same composition of each class, there are six rows of data taken randomly. Table 5 is the evaluation result.

Both test-cases give low-performance results. The accuracy and F-score of the system for both test-cases fall short of expectations. At this point, the cause of the performance issue is thought due to the underfitting. Since the distribution of the data used for training and testing is poor.

Table 4 Evaluation of the first test-case

Accuracy	Recall	Precision	F-score
0.33	0.378	0.407	0.390

Table 5 Evaluation of the second test-case

Accuracy	Recall	Precision	F-score
0.3	0.318	0.215	0.257

Table 6 Evaluation of third test-case

Accuracy	Recall	Precision	F-score
0.83	0.89	0.80	0.843

Table 7 Evaluation of forth test-case

Accuracy	Recall	Precision	F-score
0.72	0.734	0.723	0.728

4.1 Underfitting

The suspected causes are the lack of data and variance in the data. Thus, in this step, there are two additional test-cases conducted to prove the underfitting. The third test-case uses 60 rows of generated data with the proportion of training data and testing data: 70 and 30%. Here, the test uses 18 data for the testing with only two elective courses included for the classification: Cyber Security and SAP. Table 6 gives the performance results of the test.

The forth test-case uses another 60 rows of generated data with the proportion of training data and testing data: 70 and 30%. The total number of testing data is also 18 data. However, there are three elective courses included for the classification: Cyber Security, Applied Computer Networking, and SAP. The performance results are given in Table 7.

Based on these two additional tests, the reason for the low system performance is proven to be caused by the underfitting. When given a sufficient number of data with an adequate level of variances of the data, and proportionality of the data to the number of classes for classification, the Unirecommend system is able to deliver high-performance results.

5 Discussion

The Unirecommend decision support system using the Naive Bayes Classifier has been successfully implemented. The system consists of four features: recommend, score recommendation, save recommendation, and my recommendation. The evaluation based on the initial two test results gives an accuracy of 0.30 and 0.33, recall of 0.318 and 0.378, precision of 0.215 and 0.407, and F-score of 0.257 and 0.390. These performance results are considered low due to the underfitting.

The underfitting of the model is caused by the insufficient number of data used for training and testing the classifier. In addition to that, the data variance is poor. Another two tests (third and forth test-case) are conducted using a new generated data and lesser classes for the classification. These two tests show good performance

results of the model. The classification for two classes test gives an accuracy of 0.83, recall of 0.89, precision of 0.80, and F-score of 0.843. The classification for three classes test gives an accuracy of 0.72, recall of 0.734, precision of 0.723, and F-score of 0.728.

This study still has limitations, so it can be developed in future research. Here are suggestions for further research.

1. Other classification methods such as Winnowing [11] and k-NN [12] could be added to improve accuracy and F-score of the system.
2. Optimization algorithm (PSO) could be added to optimize the attributes selection process.
3. Further analysis of the data could be done to acquire the fit model for the system.

References

1. Tripathi, K.P.: Decision support system is a tool for making better decisions in the organization. Indian J. Comput. Sci. Eng. (IJCSE). **2**(1), 112
2. Xhemali, D., Hinde, C., Stone, R.: Nave Bayes vs decision tree vs neural network in the classification of training web pages. J. Comput. Sci. Issues **4**(1) (2009). ISSN 1674-0784
3. Bonczek, R., Holsapple, C., Whinston, A.: The evolving roles models in decision support system. Decis. Sci. **11**(2), 89–95 (1980)
4. Patil, T., Sherekar, S.: Performances analysis of Nave Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**(2), 256–261 (2013)
5. Rish, I.: An Empirical Study of the Naive Bayes Classifier. T.J. Watson Research Center, New York (2001)
6. Narkhede, S.: Understanding Confusion Matrix Towards Data Science. Towards Data Science. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. Cited 19 Oct 2018 (2018)
7. Manliguez: Generalized Confusion Matrix for Multiple Classes. https://www.researchgate.net/publication/310799885_Generalized_Confusion_Matrix_for_Multiple_Classes. Cited 13 July 2018 (2016)
8. Bramer, M.: Principles of Data Mining. Springer, London (2007). https://doi.org/10.1007/978-1-4471-7307-6
9. Manning, C.D., Raghavan, P., Schtze, H.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009)
10. Gorunescu, F.: Data mining: concepts, models and techniques. Springer, New York (2011). https://doi.org/10.13140/RG.2.2.31150.51523
11. Hasan, E.G., Wicaksana, A., Hansun, S.: The Implementation of Winnowing Algorithm for Plagiarism Detection in Moodle-Based E-learning (2018). https://doi.org/10.1109/ICIS.2018.8466429
12. Iswari, N.M.S., Wella, Ranny: Fish Freshness Classification Method Based on Fish Image Using k-Nearest Neighbor (2017). https://doi.org/10.1109/CONMEDIA.2017.8266036

Multilevel Video Access Control Mechanism Using Low Power Based Audio Watermarking

Sunyoung Choi, Youngmo Kim and Ung-Mo Kim

Abstract As the number of violent crimes and incidents worldwide has soared, market demand for intelligent video surveillance solutions is gradually increasing as a solution for social safety, and technology development for next-generation video solutions is underway (SME Technology Roadmap, [1]). However, existing security video transmission solutions can be consistently generated without rating authority during creation of video, resulting in problems that are shared without any control by each agency, department, or person. At this time, leakage of security video can cause serious damage to the public institutions concerned. Therefore, it is necessary to set access authority according to the intention of an original video sender and it should be transmitted inserting user's authority. The watermark information should be extracted and played in real time or in VOD according to their authority. At this time, used watermark was redesigned to improve energy efficiency. As such this paper we propose a multilevel video access control mechanism to prevent leakage of personal information in security video.

Keywords Information leakage · Transmission solution · Access authority · Low-power based watermark · Multilevel video access

1 Introduction

Worldwide, demand for video surveillance systems is on the rise. In Korea, the mandatory installation of video surveillance systems at underground parking lots has shown an average annual growth rate of more than 15% since 2000. The growth

S. Choi · U.-M. Kim
College of Software, Sungkyunkwan University, Suwon, Republic of Korea
e-mail: choisunyoung1@gmail.com

U.-M. Kim
e-mail: ukim@skku.ac.kr

Y. Kim (✉)
Department of Computer Science and Engineering, Soongsil University, Seoul, Republic of Korea
e-mail: ymkim828@ssu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_8

rate is expected to continue due to the increase of unattended stores at banks and the expansion of applications such as the strengthening of unmanned security systems, factory automation and building automation systems. As the number of cases in which the collected videos are used as legal evidence increases, the demand for the forensics technology based on video security technologies is expected to increase to prevent and respond to terrorism, violent crime. However, the development of video security technologies is required to address adverse functions such as invasion of privacy [1]. Combined with infrastructure through online, the Korean video surveillance market has evolved to various network video monitoring markets through internet networks rather than video surveillance market values that were used in previous closed networks.

However, the standardization of protocols and codecs is not prepared, and compatibility are a barrier to development. The standards required in disaster safety management, such as firefighting and police, which make up a large category as a source of demand for video surveillance values in Korea, are currently creating the standards required by the domestic industrial environment. As the market matures, the standards are being raised, and the security needs have emerged in recent years, and related laws, certifications, and standardization are expected to take place. In other countries, the industrial environment in the field of video surveillance value is not much different, but the difference is that the market for video surveillance value through local storage media is valid, unlike the domestic environment where networks are important. In addition, demand for B2C markets is on the rise, since the usage of video devices in mobile environments is higher than video surveillance markets such as drones and wearable cameras [1].

In this paper, we intend to improve the problem of existing video security system so that the video can be blocked from being leaked to the outside world, thereby implementing a multilevel video access system that allows only authenticated people can play it. Among the existing video surveillance systems, watermarking scheme is a technology that prevents illegal copying by inserting copyright information into the generated multimedia contents [2, 3]. However, the proposed mechanism allows content creators to access the content by giving the desired users equal rights, differentiating existing video security from the system. Therefore, this technology development allows the sender to encode receiver's authority that is given by sender as watermarking information. Receivers can play the video corresponding to authority. Accordingly, the security can be enhanced by preventing the outflow of video through unspecified users who do not have access to the system. This paper is organized as follows. Section 2 describes the relevant technology, and Sect. 3 describes proposed mechanism in this paper. Section 4 performs the proposed system validation through implementation and concludes in Sect. 5.

2 Related Research

2.1 Mobile Multimedia Transmission Technology

N-Screen technology has evolved from AT&T's '3-Screen' service, which allows users to synchronize content by connecting TV, PC and Mobile devices to the Internet. The N-Screen Service is a smart TV era multimedia service that provides multiple types of content and services to various wired and wireless terminals owned by individuals, as well as the features of two-way convergence, personalization and intelligence of these broadcasting services [4]. The Multimedia Delivery Solution [5] shown in Fig. 1 is a solution that can provide N-Screen service and is a technology that can extend the proposed technique in this paper. Figure 2 is a smartphone-based on-the-spot video monitoring solution that users can monitor the scene on the spot via wireless internet such as Wi-Fi, 3G and LTE during moving in real time, anytime, anywhere [6].

2.2 Role-Based Access Control Method

Role-Based Access Control (RBAC) is a multi-user, user role-based access control model for multi-programmed environments, an approach that restricts system access to authorized users in computer security system. RBAC is as a user access model associates rights and roles, and facilitates the management of rights to roles when

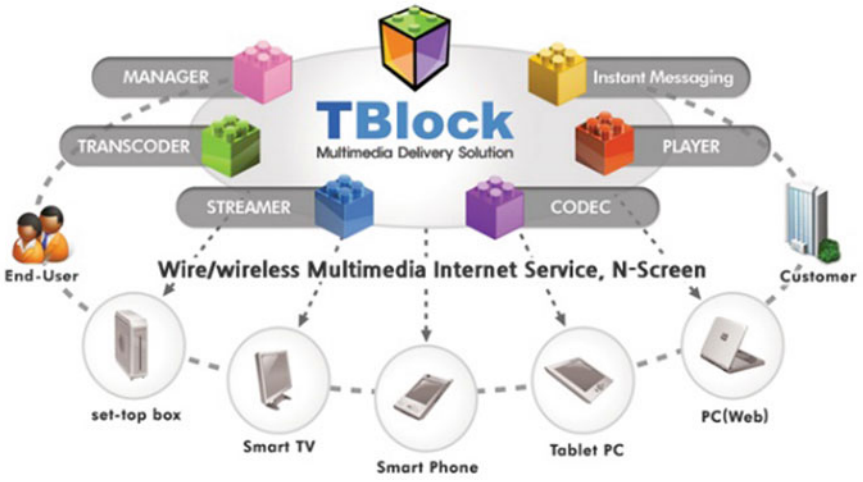


Fig. 1 Multimedia delivery solution

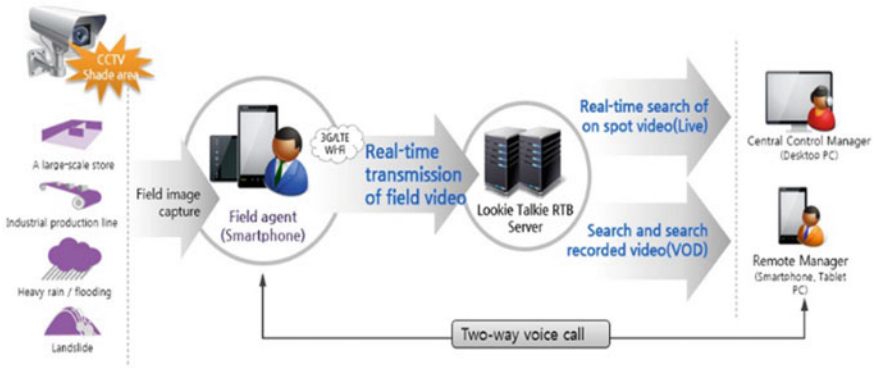


Fig. 2 Smartphone-based on-the-spot monitoring system

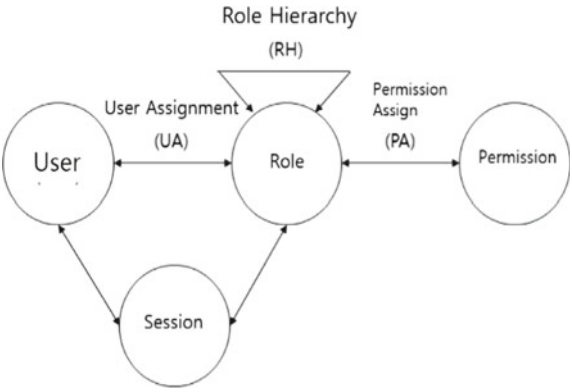


Fig. 3 Hierarchical RBAC reference model

users are assigned the appropriate roles. RBAC's layer of roles is a common way of forming roles to represent the authority and responsibilities within an organization in a partial order. The role layer is represented by a lattice structure, which is the most common way to construct a role hierarchically because it represents authority and obligations in a line. The characteristics of the role layer are that the higher role inherits the rights of the lower role and the higher role can be delegated [7, 8]. Figure 3 presents hierarchical RBAC reference model.

2.3 Audio Watermarking Technology

Watermarking technology is a technology that inserts and extracts additional data such as copyright information into multimedia contents such as images, audios,

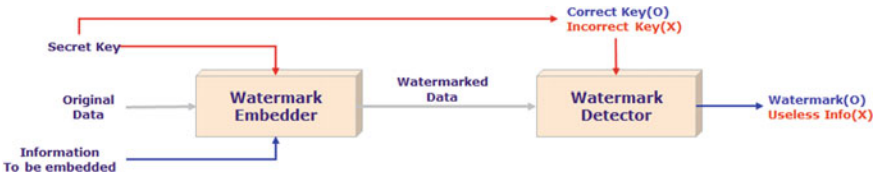


Fig. 4 Audio watermarking system

videos and text. In the event of a copyright dispute in a content, the function to resolve a dispute by extracting the inserted copyright information can be used in the copyright. Audio watermarking technology is applied to compressed audio files such as mp3 and usually consists of element technology that inserts and extracts watermark data. For audio watermarking, there is a way to insert into space domain, frequency domain, and compressed domain, and the insertion method is determined by the nature of the application field [9, 10] (Fig. 4).

2.4 Problems with Existing Systems

Although the importance of video surveillance services over the network is increasing and solutions are increasing, camera and video transmission connected over the network can cause malicious attacks and data leakage problems from the outside as the existing closed video transmission system changes from transmission over the public network. Also, because music content files like mp3 exist in a compressed format, overloading may occur when encoding and decoding. In case of each agency's own business order, it is urgent to supplement the situation by causing a delay in the rapid sharing of emergency footage of the sites held by the public organizations in case of a national disaster. This raised the need to set permissions according to access authority. Therefore, a security system is needed to secure potentially privacy-infringing video from outside attackers, and when individuals or government agencies request access to surveillance footage, a system development is needed to ensure that other people are not accessible except those who are given rights.

3 Proposed Mechanism

The proposed mechanism is deployed as shown in the configuration diagram below Fig. 5, and the contents are largely divided into smartphone security video apps that can be filmed and played, and a secure video transmission management server

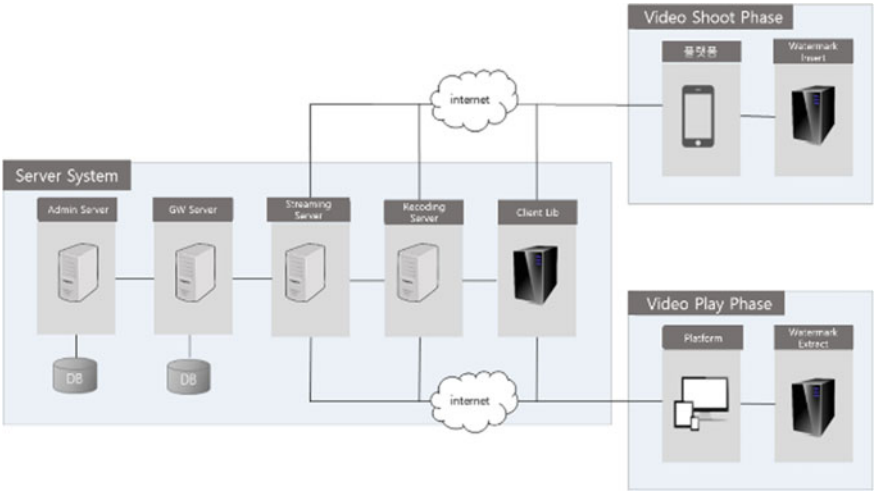


Fig. 5 Multilevel video access control system configuration plot

system that performs access control to the videos. This system is provided as an app by default and is based on N-SCREEN service. This solution is a system that inserts and transmits information using low-power-based audio watermarking techniques when transmitting videos taken using smartphones, and decodes watermark information at the video player phase to reproduce the videos in real time or in VOD according to authority (Fig. 6).

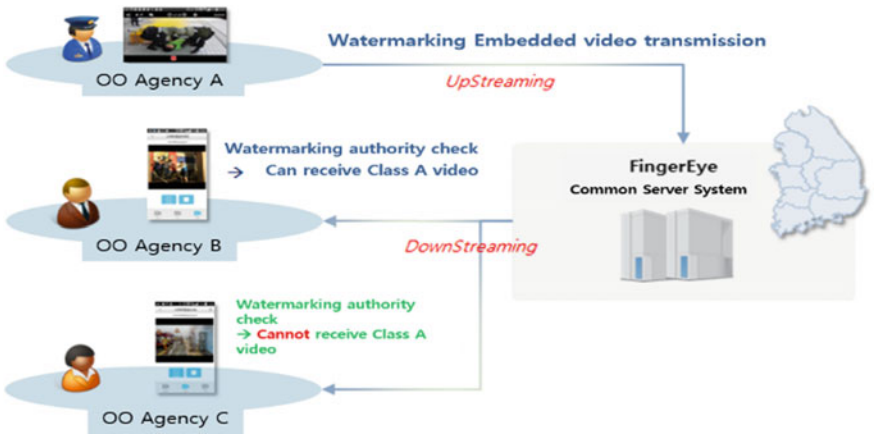


Fig. 6 Concept of video transmission/reception system based on video authority

3.1 Security Video Transfer Management Server System Design

The server system consists of the Admin Server, GW Server, Streaming Server, Recording Server, and Client Lib. The Admin Server stores and manages information such as the ID, password, agency, task, and authority rating entered by the user when he or she joined the account.

RBAC-Based Video Access Control Management

For a multi-level video access control system that specifies the right of the function to play security videos and prevents them from being used other than the specified authority, the server administrator determines the hierarchy of authority according to the agency and role of the users who sign up for membership and store it in the DB. And then, it specifies the rating criteria to the video creator and allows the user to choose who will be allowed access when uploading videos. When a video request is made in the play phase, the task information is analyzed with the agency of the device's registrar and video is provided only if they match the allowed privilege rating. The object configuration of the role-based access control model is divided into users, roles, permission, and sessions, and the definitions for each configuration utilized in a secure video system is shown in Table 1.

Watermark-Based Video Access Authorization Server

The steps for processing video authentication in the terminal are as follows. If Announcer, video transmission terminal, of the video creator is authenticated (ID/Password) by the server and streams the video, it requests to create a live video session to the Service Server. Upon receiving the request, the Service Server passes information about the Media Server URL and connection OTP and the rights (AuthlLevelID) of the video to the transport terminal. The Player that is a viewer terminal requests creation of a live video session to Service Server in order to obtain server authentication information such as Media Server URL and OTP. And then, it accesses

Table 1 Reference model components

Object	Describe
User	Someone who requests access to multi-level video access control services
Role	Divide layers by their organization and task to give authLevelIDs to each layer
Permission	Define permission levels based on user information Control access based on the authLevelID assigned in the role
Session	Mapping a user with an active subset of the assigned roles

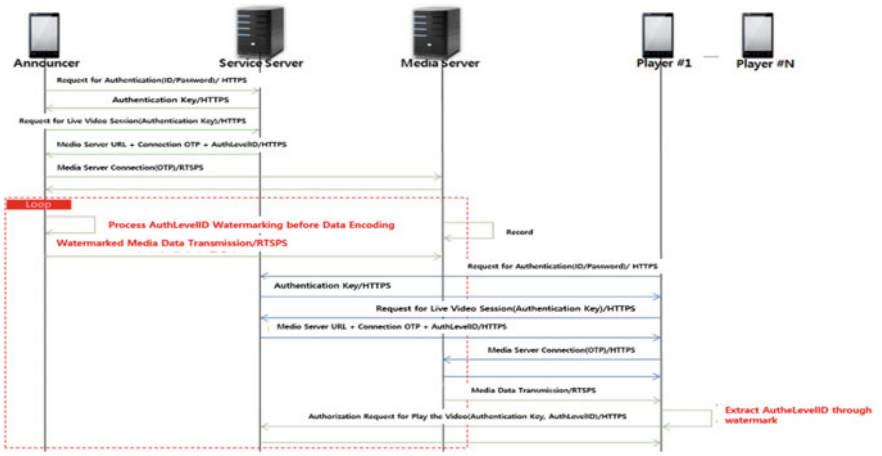


Fig. 7 Diagram for video/certification processing within announcer

Media Server through the information received from the Service Server. Depending on your user account, you can view list of videos stored on the Media Server and select video. Once the selected video is transferred, the watermark can be extracted and compared to the authority information (AuthLevelID) and the video can play once certification is completed (Fig. 7).

GW Server

The GW server is an authority agent, which provides multi-level access control based on it, when the viewer's video requests after the administrator assigns access rating by agency and task and stores it in the DB. Streaming Server, Recording Server, and Client Lib are video management systems that receive and transmit watermarked video containing rights rating information from a secure video sender to the requesting players. It uses RTP protocol for real-time transmission.

3.2 Smartphone Secure Video App Design

Smart security video apps are used to upload videos taken by users or to play them on a device. Video makers shoot disaster video through the app, specify access rights information, and upload them to the server. At the same time as the upload, watermark containing the maker information and authority information is inserted in real time.

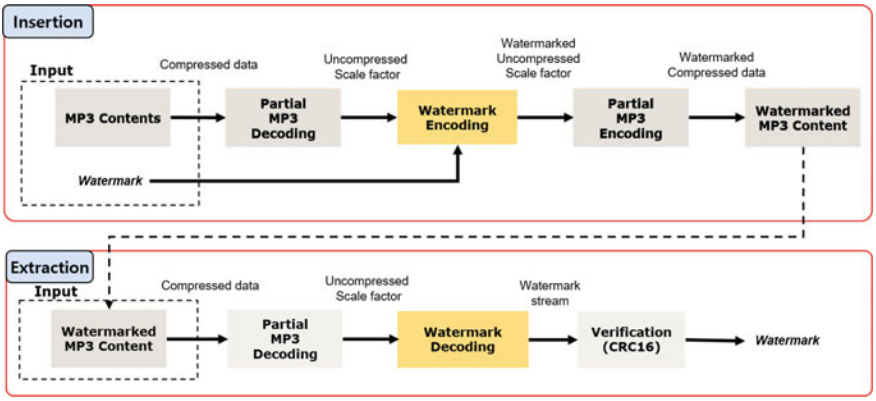


Fig. 8 Audio watermark flow chart

Low-Power-Based Watermarking Technology

Watermark information is inserted and extracted through partial decoding and encoding of audio formats in mp3 to reduce overload during insertion and extraction of watermarks. In addition, for energy efficiency refactoring was applied in the watermarking technology. This audio watermarking technology inserts watermark using scale-factor during the mp3 audio format [9]. Further, the MP3 partial decoder and encoder for insertion and extraction were developed and applied by ourselves, and the usage information ID consists of 64 bits of ASCII 8 characters. It is validated watermark information extracted using CRC. Figure 8 shows the flowchart of the audio watermarking.

Insertion of the generated watermark inserts in the scale factor extracted by the partial mp3 decoder according to the following formula rules.

- Insert 1 watermark 1 bit per frame of the mp3 decoder.
- Adjust the total sum of the scale factor of the frame to be odd or even relative to the watermark bit, so that the coefficient value does not exceed 7.
- The insertion is repeated in proportion to the length of the audio frame to be inserted. For example, if there are 800 frames of mp3 audio, the watermark (80 bits) is repeated 10 times.

The watermark insertion pseudo-code according to the above rules is shown below.

– S_i: Sum of Scale factor values
– L_i: Last index value of Scale factor

$$L_i = \begin{cases} L_i + 1 \text{ If } S_i \text{ is odd, } W_i \text{ is zero and } L_i \neq 7 \\ L_i - 1 \text{ If } S_i \text{ is odd, } W_i \text{ is zero and } L_i = 7 \\ L_i \quad \text{If } S_i \text{ is odd, } W_i \text{ is one} \\ L_i + 1 \text{ If } S_i \text{ is even, } W_i \text{ is one and } L_i = 7 \\ L_i - 1 \text{ If } S_i \text{ is even, } W_i \text{ is one and } L_i = 7 \end{cases}$$

Some of contents are decoded to extract the inserted watermark, then the masked watermark value is extracted using the scale factor value. If the sum (S_i) of the extracted scale factor is even, mask it to 0, if odd, mask it to 1, and extract the watermark bit (W_i) value

$$w_i' = \begin{cases} 0 \; If \; S_i \; is \; even \\ 1 \; If \; S_i \; is \; odd \end{cases}$$

S_i : *Sum of Scale factor values*

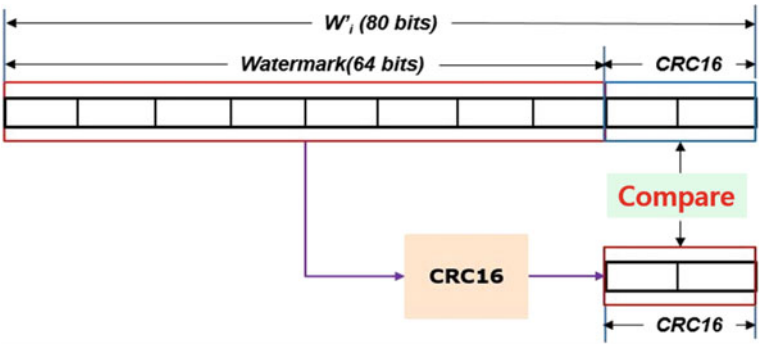


Fig. 9 Extract watermark structure

The CRC16 for 64 bits of extracted watermark (80 bits) is recalculated and matched with the extracted CRC16 value, and is considered a watermark extraction success only if all 16 bits are matched [9–11] (Fig. 9).

Development of Watermark-Based Access Control

Multilevel access control system of watermark-based security footage is shown in Fig. 10. At Announcer State, it accesses the app, disconnects the audio of the video taken with the camera, and inserts the watermark according to the audio watermarking algorithm described above. Video files that are re-compressed with watermark information are stored on the Media Server via the transmitter. If the Player wants to receive and play video, it will receive and decode it from the Media Server in real time through the receiver. The inserted watermark information is extracted and compared it with the Authkey and AuthLevel IDs stored in the Service Server. Rendering video only when the verification phase is complete and the information matches, overlaying the screen.

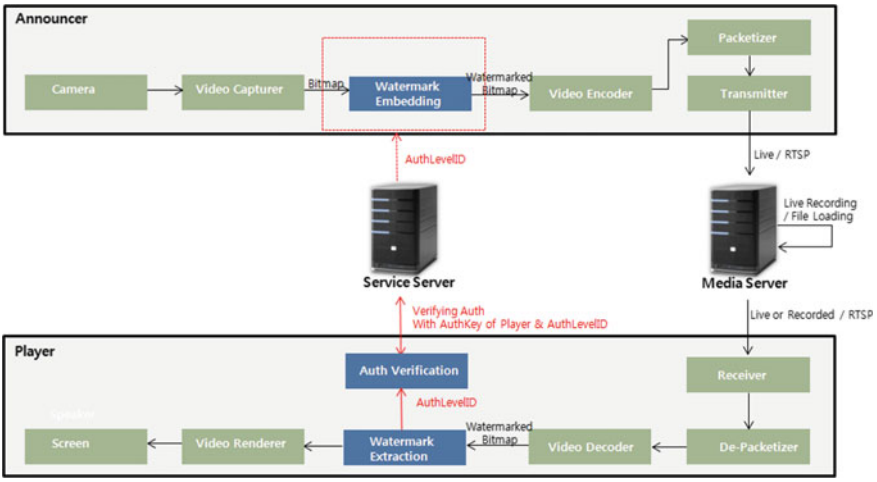


Fig. 10 Multilevel access control system of watermark-based security footage

4 Validation

On the Web Permissions Management screen of the Admin Server system, it verified that the user's access to video by account can be registered/modified/deleted, and that a user's query based on the video access authority is possible, and can log in as a registered user via the mobile device. The test environment and test target are shown in Fig. 11 and Table 2.

It checked video access rights information by user account has been registered/modified/deleted/user lookup function enabled. A new user has registered and logged in on a mobile device. The user's authentication information like Fig. 13 was verified through the server's log file (Figs. 12 and 14).

Each user logged in Server from the mobile device and accessed the video of the server that meets the authority to ensure that the video is played. When users logged in on mobile devices, they checked if they could play video files that match access rights, and if they were not authorized, they found that video files could not be played. Also, we measured the speed it takes to insert/extract the watermark and

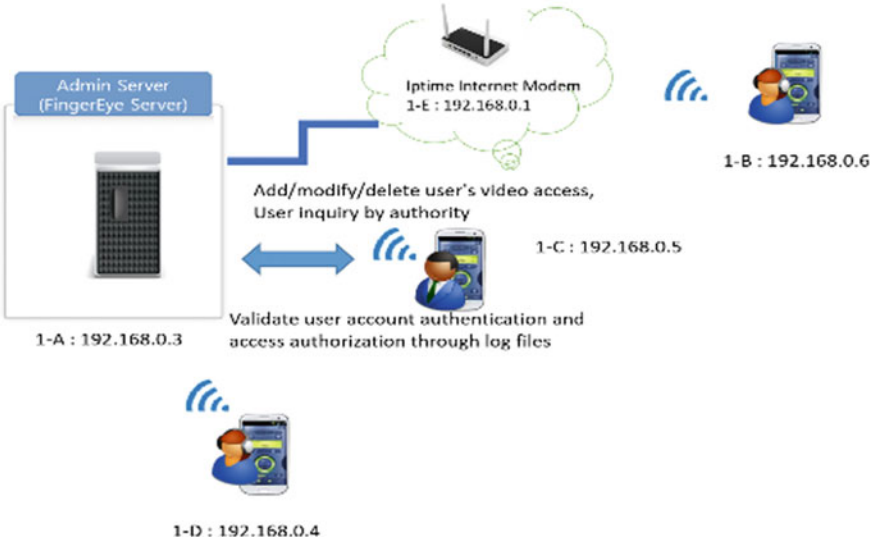


Fig. 11 Test environment for rights authentication

Table 2 Test target

Category	Security video transfer management server detail	Smartphone secure video app detail	Executable file for checking watermarking module functionality
Hardware	Intel(R) Core(TM)i5 CPU M520 @2.40 GHz	Samsung Galaxy S3(SHV-E210S) 2	Intel(R) Core(TM) i5-2520M CPU @2.50 GHz
	4 GB RAM	LG AKA(LG-F520L)	8.0 GB RAM
		iptime Internet Modem(Wifi) 1	
Software	DB: Mysql 5.5	Fingereye App 1.0.3 (proposed implementation App)	Watermarker.exe(1.0)
	Was: Apache Tomcat 6.0.37		
	Spring 3.02 framework, Eclipse, Visual Studio		
OS	Windows 7 Home Premium K, 32bit	Android version 4.4.2	Windows 7 Professional K

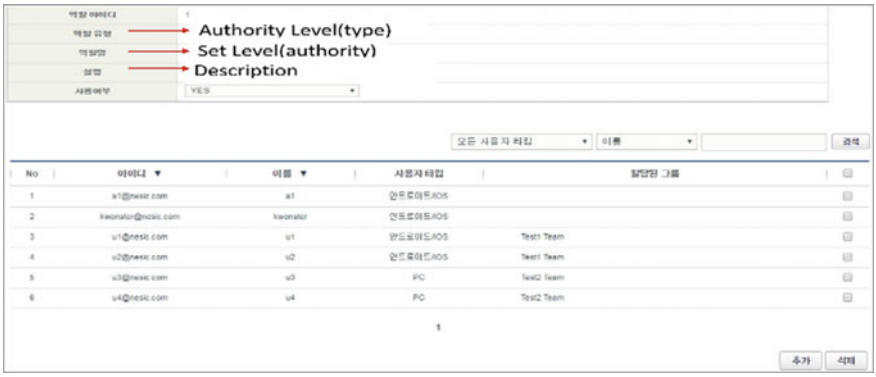


Fig. 12 User permissions settings UI\

Fig. 13 Admin server's log file

calculated the consumed time per Mbyte and checked it that time was within 2 s. Tables 4 and 5 shows the results of 10 insertions and extractions (Table 3).

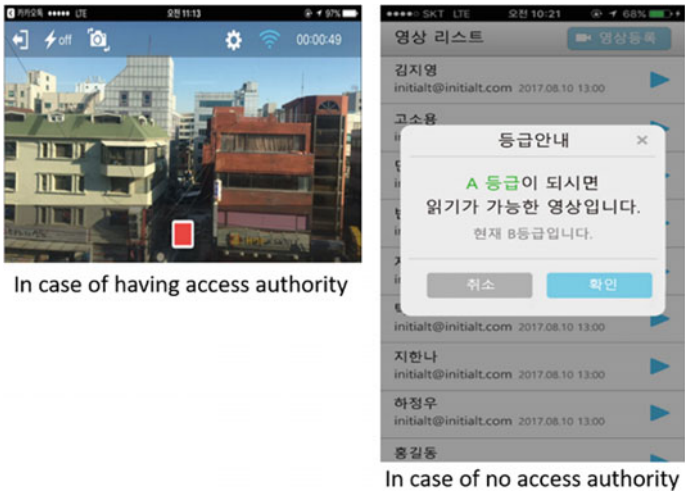


Fig. 14 Check if video is played according to access authority

Table 3 Playable ratings and results by access level

Video access level (authority)	Playable rating	Test result
A	A	A
B	A, B	A, B
C	A, B, C	A, B, C

Table 4 The average speed time for watermark insertion

Test order	Size (MB)	Start time (s)	End time (s)	Consumed time (s)	Speed (consumed time/MB)
4-1	9.38	02:31:49.235	02:32:01.368	12.133	1.293
4-2	8.92	02:35:55.708	02:36:02.786	7.078	0.793
4-3	13.6	02:38:13.086	02:38:21.156	8.07	0.593
4-4	3.43	02:39:44.991	02:39:53.071	8.08	2.356
4-5	8.41	02:40:53.121	02:40:58.175	5.054	0.600
4-6	9.12	02:41:55.077	02:42:03.155	8.078	0.886
4-7	9.36	02:43:43.952	02:43:48.999	5.047	0.539
4-8	9.26	02:44:54.271	02:44:59.324	5.053	0.546
4-9	9.27	02:46:52.862	02:46:59.930	7.068	0.762
4-10	8.74	02:49:35.616	02:49:46.736	11.12	1.272
Average speed					0.964

Table 5 The average speed time for watermark extraction

Test order	Size (MB)	Start time (s)	End time (s)	Consumed time (s)	Speed (consumed time/MB)
5-1	9.38	03:48:32.491	03:48:40.512	8.021	0.856
5-2	8.92	03:48:40.512	03:49:27.661	47.149	5.286
5-3	13.6	03:49:51.904	03:49:57.920	6.016	0.443
5-4	3.43	03:51:47.791	03:51:52.791	5	1.458
5-5	8.41	03:52:11.132	03:52:16.226	5.094	0.606
5-6	9.12	03:52:16.226	03:52:41.818	25.592	2.806
5-7	9.36	03:52:59.659	03:53:05.675	6.016	0.643
5-8	9.26	03:53:22.112	03:53:28.127	6.015	0.650
5-9	9.27	03:53:47.309	03:53:52.325	5.016	0.541
5-10	8.74	03:54:11.927	03:54:17.953	6.026	0.690
Average speed					0.155

5 Conclusion

In this paper, we proposed multilevel video access control mechanism using low-power-based audio watermarking in order to solve the problem of reckless playback and sharing of existing security video transmission solutions. This mechanism is inserted into the video using watermark techniques redesigned as refactoring for energy efficiency and it can automatically decode watermark information by authority. After decoding, it receives videos and can play corresponding to authority. It is not only enhancement of security by preventing video leakage through unspecified users without viewer rights, but also allowing video information from each public institution to be shared in emergency situations such as war/disaster/recovery/emergency. It is expected that applying this technology to the national disaster system will improve the ability to respond early in case of emergency and become an efficient national disaster system through the rapid integration command system.

References

1. SME Technology Roadmap: http://smroadmap.smtech.go.kr/0201/search/m_code/S10/id/1215. Technology Roadmap for SME 2015–2017
2. Kim, Y.: Black box, service providing method for vehicles enhanced security based on watermarking technique. KOREA PATENT/1020120060072
3. Kim, K., Kim, M., Yim, H., Kim, D., Yu, J.: Video transmitting system and method, and video play terminal and method. KOREA PATENT/1020110031376
4. Yoon, C., Jeon, S., Lee, H., Lee, K., Lee, H., Ryu, W.: Classification of N-screen service and standardization. Commun. Korean Inst. Inf. Sci. Eng. **29**(7), 23–31 (2011)

5. initialT Co., Ltd.: Multimedia delivery solution TBBLOCK. http://initialt.com/newhome/tblock.html
6. initialT Co., Ltd.: Sharing of real-time multimedia information based on smartphone. http://initialt.com/newhome/lookietalkiertb.html
7. Cho, KiCheon, Shin, MoonSun, Ryu, KunHo: Structure of role hierarchy for user-level delegation and inheritance in role-based access control. The Korea Soc. Inf. Technol. Appl. **05**, 107–110 (2001)
8. Lim, I., Kwon, D., Kim, H., An, D., Jus, H.: Design of user access control model based on RBAC for service mash-up system. J. KIISE **12**, 879–881 (2016)
9. Fallahpour, Mehdi, Shirmohammadi, Shervin, Semsarzadeh, Mehdi, Zhao, Jiying: Tampering detection in compressed digital video using watermarking. IEEE Trans. Instrum. Meas. **63**, 1057–1072 (2014)
10. Kirovski, D., Malvar, H.S.: Spread-spectrum watermarking of audio signals. IEEE Trans. Signal Process. **51**(4)
11. Choi, S., Koo, J., Kim, U.-M., Kim, Y.: A study on multilevel video access control method using low power based audio watermarking. EEECS 2018 (ISSN 2466-152X) http://iaser.org/Vol-6/, pp 24–28 (2018)

A Study of Persona Research on Domestic Music Applications

HaeKyung Chung and JangHyok Ko

Abstract If you look at the download ranking of users in the domestic Google Play music and audio parts, you can see that the streaming app is popular. This is because you can easily listen to the music you want and the recommended music for each category. In this study, we survey Melon, Genie, and Bugs which are the most users through the questionnaires of college students. In this study, firstly literature study was conducted to understand overall understanding of mobile music service, and completed persona through survey methods such as surveys and in-depth interviews. The user's goal was simple, easy to operate the app easily, and wanted to listen to music stably and seamlessly. Needs wanted diversity of content and low price plan. I would also like to make it easier to find non-mainstream music, and I wanted to decide whether to use the additional services in the unnecessary apps by selecting ON/OFF. The interface design aspect required a simple yet intuitive design. Since most of the users are students, they were sensitive to the fare. Nevertheless, I wanted to enjoy a lot of music, and I wanted to use less data when streaming. I wanted to concentrate more on music than on additional services, and all these needs in one way was that users wanted a simple, lightweight music application.

Keywords Music apps · Usability test · Persona

1 Introduction

The development of domestic IT technology and the internet usage environment have developed rapidly, and the way in which music contents are consumed has changed a lot due to the popularization of smart phones owned by individuals. In particular, as

H. Chung (✉)
Department of Visual Communication and Media Design,
Konkuk University, Chungju-si, Republic of Korea
e-mail: jangmi44@gmail.com

J. Ko
Division of Computer and Mechatronics, Sahmyook University, Seoul, Republic of Korea
e-mail: janghyokko@syu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_9

LTE (wireless communication technology) became popular and personalized media consumption increased, music contents became more representative contents than other contents genre. Music contents, which are small in file size and easily accessible through personal devices, are easy to share, link, and utilize in the music industry and related industries. In the past, if music could only be played on a specific device owned by an individual through storage, music consumers can now listen to music anywhere, anytime in a ubiquitous environment. Currently, consumers are paying a certain amount and listening to music through a streaming service instead of downloading and owning a monthly fee [1].

If you look at the download ranking of users in the domestic Google Play music and audio parts, you can see that the streaming app is popular. This is because you can easily listen to your favorite music and recommended music for each category. In this study, we survey Melon, Genie, and Bugs, which are the most users through the questionnaires of college students. First of all, the purpose of this research is to find out the directions and ways to improve the usability of mobile music service based on the research results.

In this study, we used Melon, Genie, and Bugs, which are related to Naver music application comparison query and cumulative downloading, among the domestic music applications. As a research method, first of all, literature study was conducted to understand overall understanding of mobile music service, and persona was completed through survey methods such as questionnaires and in-depth interviews.

2 Related Works

2.1 Usability Evaluation Methodology

2.1.1 Types of Usability Evaluation Methods

There are many ways to assess usability, including ethnographic studies, participatory design, focus group research, surveys, walk-throughs, card sorting methods, paper prototyping. These methods are quantitative or qualitative. But in common, all these forms of evaluation help designers and developers to develop services and products that can bring users closer to their expectations and their goals [2].

In this study, we made an affinity diagram through surveys and in-depth interviews among various methodologies and completed the persona with the extracted keywords.

Rubin and Chisnell (2008) argue that available products should be useful, efficient, effective, satisfying, learnable, and accessible [3].

In a similar vein, Barnum (2011) states that tools must be easy to learn, easy to use, intuitive, and fun. In addition, the International Organization for Standardization (9241-11) defines the definition of usability as “the extent to which a user achieves

certain goals with effectiveness, efficiency and satisfaction when using a product in a particular environment of use" [3].

Rubin and Chisnell argue that the most important aspect of usability definition is that users are not frustrated while using a product or service and can naturally perform the task they want without any questions.

Barnum (2011) argues that usability should not be seen. In other words, the concept is that the built-in usability of the product is appropriate for the user, and that the user is not motivated by the product's will [4].

2.2 Music Application Service

2.2.1 Melon

Melon Music started service in November 2004, and YBM Seoul Records, which was the number one distributor of music in Korea, is the beginning. Currently, it is a music service operated by Loen Entertainment, a subsidiary of Cacao.

Owned by Loan Entertainment, which was sold to KaKao in January 2016, it was originally owned by SK Telecom but was handed over to its subsidiary, Loen. In 2013, it sold 15% of its stake in Star Invest Holdings Limit (SIH), but did not sell the yen at the time of sale, and when the acquisition of 2016, KaKao absorbed the stake in the company as SK, It does not mean that SK has been disconnected. Even after 2013, you still get a 50% discount on T-membership at regular payments. Since the existing Melon phones are still available and there are membership benefits, it does not change the relationship of win-win regardless of the sale. However, the way of running under the KaKao should be watched [5].

Melon provides a graph of the real-time rankings from the first to the third, showing a high real-time uptick, and the graph rises above the limit, also referred to as 'pierced roof'. Note that the real-time ranking is relative, so it's hard to imagine that a lot of download music or number of streaming music.

The latest music and real-time charts provide services, clever recommendations to know my tastes, and a wealth of information to keep you updated on the latest news from your favorite artists.

The most representative of services.
The first [Melon chart] provides various charts by synthesis, sound source, album, and age.
The second [Melon DJ] provides DJ playlists with high-end selection of Power DJ's by listening and listening with tags.
The third [Melon Radio] is based on my Melon activity history, analyzing the taste of music, recommending music or recommending a music star DJ.
The fourth [Melon TV] music video, broadcasting, as well as all the images of the artist provides.

The fifth [MY MUSIC] allows me to gather various activities that I have done in Melon at a glance, such as my music usage patterns, fans' artists and friends.
The sixth [ForU] analyzes my interest in evil and music DNA to recommend music that is right for me.
The seventh [feed] allows you to see the latest news from artists such as album releases, concerts, and live broadcasts.
The eighth [music search] allows you to search smart music simply by playing music.

2.2.2 Genie

The predecessor of KT Music, which has a Genie, is Shinseong Technology Research Center, a venture company established in 1991 for technology research and development and semiconductor manufacturing. It was listed on the KOSDAQ in 2000 after the company changed its name to BlueCode Technology in 1999. It has grown by winning the Presidential Prize of Venture Company in 2001 [6].

It is currently in the second place on the domestic music source site, and in fact, it is said that the ranking itself is more popular than Melon nowadays. In the case of Melon, it is hard to get the rank of the top 100 in the top rankings because the streaming rankings are frequently shaken while the male idol group fans are occupying a lot of positions. On the contrary, the Genie music is less.

If you look at the features of the service, first of all, there is [Home menu DIY function], you can move to the main menu from the home screen, you can change the position of the menu as you like, and you can freely adjust the size of the menu like widget.

The second [My Style] recommends custom music based on your listening history.

There is a third [recommendation selection], which enables a sensible selection with various recommendation lists tailored to the tag.

It is the first in the industry as the fourth [VR exclusive hall], and provides 360° vivid images.

The fifth [Genie Life] offers a variety of functions that can be combined with music to suit the user's life cycle [7].

2.2.3 Bugs

Bugs Music is the first music service in Korea released in 1999. It is a sound service provided by NHN Entertainment's subsidiary, Bugs. In the early days of the business, the company was forced to go bankrupt due to the music business. However, Eins Digital, who operated Jukon, changed its name to NeowizBugs after receiving the investment of Neowiz and changed its name to Neowizb internet, and Jukon as a site under the Bugs brand. In other words, the old Bugs were closed, and the current Bugs is the old form of the jukebox. The reason why the Jukon brand has been sold is because of the higher recognition of the Bugs than the Jukon in the B2C market.

In the B2B music market, Jukon topped Melon in the deal, but in the B2C market, it is in second place after Melon.

The features of the service
The first [Music 4U] is the most representative, recommending music with intelligent personalization based on big data.
The second [radio] allows you to listen to the theme, the atmosphere and listen to your favorite from the radio menu.
It is a service to share the music with the service that can open the album selected by the third [Music PD] to other users.
You can view and play the chart directly on the phone screen without opening the app with the fourth widget.
The fifth [Playback mode selection] allows you to play all the songs instantly, and you can select and play back your personal preferences [8].

A brand is helpful in maintaining customer loyalty and the long-term operation of a company. High service quality will increase customer satisfaction and make the users more willing to recommend the website to other people and generate the effect of public praise marketing [9].

A well-planned trial and service promotion can attract more users. In addition, in relation to membership activity and discounts, a better member discount program and member activities can enhance the loyalty of members [10].

The decision-making process of online shopping will affect the complicacy of shopping behavior. Therefore, good recommendation function will help consumers to make decisions. The most popular supporting search functions of online music stores are songs or music ranking, album category, and the introduction of the album or singer, and real time recommendations [11].

3 Empirical Study for Usability Improvement

3.1 User Survey

3.1.1 In-Depth Interview with Users

In-depth interviews were conducted with a total of six, two Melon users, two Genie music users, and two Bugs music users. The age group decided to be the most popular college student of Music App, and did not consider gender specifically [12, 13] (Table 1).

Melon Music User 1 said that he used Mnet and changed songs to Melon because he had a lot of songs. In terms of usability questions, it was difficult to find a search box. In other words, Melon has a search box with a single green line on a white background. However, in terms of the diversity of music, I have a variety of music and I was satisfied with it. The add-on is that the real-time chart is easy to see on the

Table 1 In-depth Interviewees

Melon users	22 year old college student, female
	27 year old college student, male
Genie users	25 year old college student, male
	24 year old college student, male
Bugs users	20 year old college student, female
	23 year old college student, male

home screen, but there is no need for manipulation, and it is intuitive and verifying useful information. The most uncomfortable thing and the thing that wants to be improved is that it is not possible to move at once because the process of setting the playlist is complicated.

Melon Music User 2 also noted the inconvenience of the complexity of the playlist setup process. He also said that he did not use the supplementary service function a lot and thought it was unnecessary. Also, when changing the fare system, it is divided into immediate change and change at the end of this month. However, it is easy to download music and is easy to see at the left menu.

Genie Music User 1 said that it was used as a basic option since buying a cell phone and because it was able to get a discounted rate through a partnership with KT. In terms of usability, the web is easy to use, but it is not intuitive because it is designed to be relatively hard to find in apps. For example, some of the frequently used supplementary services are becoming inconsistent even if the app suddenly grows. The additional service features that are useful are the music hug (an add-on that can listen to music in the form of a personal radio) so that you can listen to the music without looking for the music several times. What I would like to improve is that the lock screen is also in the music app, which means that the lock on the mobile phone itself and the lock on the music app are duplicated, which makes it very cumbersome to unlock.

Genie Music User 2 insisted that fan-out of the supplementary service functions seemed to have no relevance to music and that it seemed to be described above, and claimed that additional functions were unnecessary. And it is generally satisfied with the kind and diversity of music. They said that the fare was reasonable, and that there were not many inconveniences in settlement or termination.

Bugs Music User 1 said that the reason she used was reasonable prices. Additional useful services are real-time charts, music videos and themed selections. As for usability, music retrieval is easy and layout is easy, so we can easily find the elements we want to find clearly.

Bugs Music User 2 said that it was used because of the variety of music and the reasonable price system. Additional service features that you do not need are Music PD albums, music posts, and so on. Among the additional service features, Music 4U is a function that grasps my music taste and recommends music every day.

3.1.2 Survey

The purpose of the questionnaire was to investigate the user 's perception of mobile music service application and to find improvement by deriving the usability problem through users' usage of music service (Table 2).

The age of the questionnaires was all in their 20s and college students, and the questionnaires were conducted as online surveys. It was conducted for one week from April 19, 2017 to April 26, 2017. The total number of respondents was 124, except for the three respondents who said they had never used a music app.

The frequency of use of mobile music service was 45, 36.5% for 3–4 times a week, 16.5% for 1–2 times a week, 2% for 1–2 times a month, of users use 3–4 times more than 80%.

The questionnaire items are used to determine whether the desired music source is searched and saved or whether the process of storing the music in the list is simple, whether the process of listening to and finding the music stored in the list is simple, whether the playback related functions such as repeated playback and shuffle are easy, Whether the features are used frequently, there are no problems with the multitasking function, the content category is satisfactory, the supplementary services are used frequently, the text size is appropriate, the icons are designed to be easy to see and understand, Asked about harmony.

- **Melon**

It is an application used by the greatest number of people in the survey respondents. Most users responded positively to the use of the application, but most of the users in the supplementary services area did not need or know features other than real-time charts and recommendations.

Some commented that the payment and cancellation process was inconvenient, and there were opinions that data management was difficult. There were comments that users who want to listen to foreign or non-mainstream music do not provide various sound sources.

- **Genie**

Most users were satisfied with the application because they had more music than other applications. Jeannie users were chosen because they were cheaper than other applications. Additional services other than the recommended selection were unnecessary or did not even recognize existence.

Student users said they would like less data while they use the apps.

Table 2 Music application users

Application users		
Melon	68 people	56%
Genie	34 people	28%
Bugs	19 people	16%
Total	121 people	100%

- **Bugs**

Most users thought the application was convenient, especially the search function. Other than the real-time charts, I did not use the add-on features well, and I did not want to spend more than 20,000 won a month on using the Bugs application.

3.1.3 Affinity Diagram

Affinity Diagram is defined as affinity, which is used to mean "collect it" in UX.

It is commonly used to mean collecting unit data that can reveal the relevance of user experience (UX) from field research.

The reason why an affinity diagram is needed is that you can discover the rules of data during a number of user-search phenomena.

Through affinity diagrams, we can find out the regularity of the data in the process of analyzing the association of various data, and model it, and discover the current problem and the improvement through the modeling. It also helps create new things. Table 3 shows, that core keywords are extracted with diversity of reserved music, low price plan, low data usage, and additional services.

In Table 3, the results showed that most of the users were sensitive to the plan because they were students and nonetheless, they were young and wanted to enjoy a variety of music according to their personality. I also wanted to use less data when streaming and wanted to focus more on music than on additional services. In addition, all of these keywords combined into one, users want a simple, lightweight music application.

3.2 User Modeling

3.2.1 Deriving the Point of Occurrence of Trouble Through Customer's Travel Guidance

The point at which the problem occurs is complex, difficult to use because there are too many add-ons that were first, unnecessary, and unknown. Secondly, there is a lack of diversity of non-mainstream songs and foreign songs. Third, fewer data usage, fourth, low-cost plans.

- Key problem 1: There are too many add-ons that are unnecessary and unknown, which interferes with complicated and important tasks (searching and listening to music).
- Key problem 2: I want to listen to various music such as minor songs, foreign songs, etc.
- Key problem 3: Too much data usage during streaming.
- Key problem 4: It is too expensive if the student does not use it regularly (Fig. 1).

Table 3 Affinity Diagram

Simple lightly enjoyable I want a music app	Variety of music	The amount of sound source provided is insufficient	Feeling lack of foreign music sources
			Each application has different types of music
		Various services for music are needed.	International music update is slow
			Old song recommendation function
	Cheap plan	The price is burdensome	Plan is not reasonable
			I feel like I'm using an app but it's expensive
			Most of the users are students
			No users want to spend more than $20 on music apps
	Data usage	Data is heavy	Use more streaming than download
			I want to be able to manually adjust data usage
	Extra service	The additional services that are mainly used are limited	Real-time chart and recommendation selection among supplementary services are most useful
			I think most users do not need the magazine function
			I have a favorite singer, but do not use pan
		A simple interface is needed	Complex interface with various additional services
			Payment is easy, but cancellation process is complex

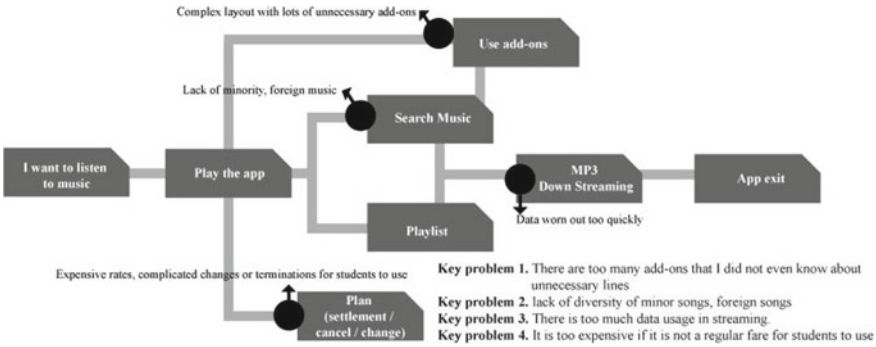


Fig. 1 Customer journey map

3.2.2 Persona

As shown in Fig. 2, Persona 1, Sim Chi-ling are 23-year-old women who use the Genie. The goal was simple, easy to operate the app easily, and wanted to listen to music stably and seamlessly. Needs wanted diversity of contents and low price plan. Her goal is to make it easier to find non-mainstream music, and not to see add-ons in

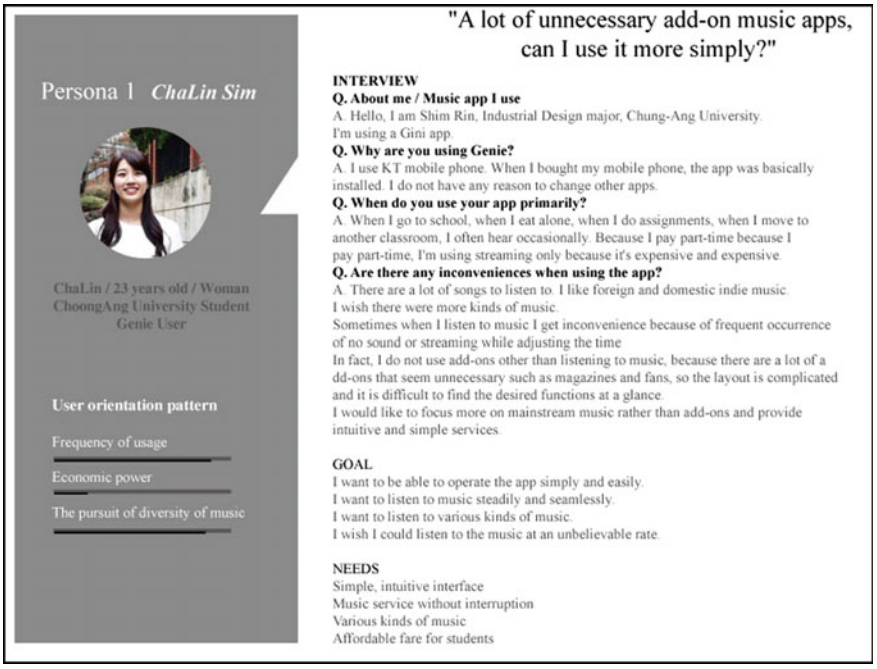


Fig. 2 Persona

unnecessary apps. And also her needs are that she would like to have a design layout selection function that includes the ability to turn on/off unnecessary add-ons and services that are immediately updated immediately upon request for minor music.

INTERVIEW

Q. About me/Music app I use
A. Hello, I am Shim Rin, Industrial Design major, Chung-Ang University. I'm using a Gini app.
Q. Why are you using Genie?
A. I use KT mobile phone. When I bought my mobile phone, the app was basically installed. I do not have any reason to change other apps.
Q. When do you use your app primarily?
A. When I go to school, when I eat alone, when I do assignments, when I move to another classroom, I often hear occasionally. Because I pay part-time because I pay part-time, I'm using streaming only because it's expensive and expensive.
Q. Are there any inconveniences when using the app?
A. There are a lot of songs to listen to. I like foreign and domestic indie music. I wish there were more kinds of music. Sometimes when I listen to music I get inconvenience because of frequent occurrence of no sound or streaming while adjusting the time.

In fact, I do not use add-ons other than listening to music, because there are a lot of add-ons that seem unnecessary such as magazines and fans, so the layout is complicated and it is difficult to find the desired functions at a glance. I would like to focus more on mainstream music rather than add-ons and provide intuitive and simple services.

GOAL

I want to be able to operate the app simply and easily.
I want to listen to music steadily and seamlessly.
I want to listen to various kinds of music.
I wish I could listen to the music at an unbelievable rate.

NEEDS

Simple, intuitive interface
Music service without interruption
Various kinds of music
Affordable fare for students

To summarize the needs of users, as mentioned in the result part of the affinity diagram, they wanted to enjoy various music on a low price plan, and wanted to decide on their own choice through the function of ON/OFF unnecessary supplementary services. Users also wanted a simple yet intuitive interface design.

4 Conclusion and Suggestions

In the past, it was possible to listen to music only on specific devices owned by individuals through storage. Recently, consumers are mainly using services to listen to music through streaming instead of owning a sound recording. In this study, we researched Melon, Genie, and Bugs, which had many queries related to Naver music application and accumulated number of downloads, and users in questionnaires.

The results show that most of the users are students, so they are sensitive to the rate plan. Nevertheless, as young as they are, they have found that they want to enjoy various music according to their personality. I also wanted to use less data when streaming and wanted to focus more on music than on additional services. In addition, all these needs are combined into one that users want a music application that is simple and light to enjoy.

The result of each application is as follows.

First, Melon music is the application used by the largest number of respondents. Most users responded positively to the use of the application, but most of the users in the supplementary services area did not need or know features other than real-time charts and recommendations.

Some commented that the payment and cancellation process was inconvenient, and there were opinions that data management was difficult. There were comments that users who want to listen to foreign or non-mainstream music do not provide various sound sources.

Second, Genie Music was satisfied that most users were satisfied with the application because of the variety of music they had compared to other applications. Also, many of the Genie users mentioned that the reason for the selection is that the fare is cheaper than other applications. Additional services other than the recommended selection were unnecessary or did not even recognize existence.

Third, most of the users thought that the application was convenient, especially the search function. Other than the real-time chart, I did not use the supplementary service function well.

The limitation of this study is that it is difficult to generalize the study results because the respondents are limited to the twenties college students. Nonetheless, the age at which people use music applications the most is worth researching because they are in college, especially in their twenties, and they have been able to draw implications for research topics. In the future, it will be necessary to study respondents for various ages.

References

1. Kim, J.B., et al.: A study on the influence of free music streaming service on the music industry. Korea Creat. Content Agency **3**(4), 12–22 (2015)
2. Gilmore, J.S., Engelbrecht, H.A.: A survey of state persistency in peer-to-peer massively multiplayer online games. IEEE Trans. Parallel Distrib. Syst. **23**(5), 818–834 (2012)

3. Färber, J.: Network game traffic modelling. In: Proceedings of the 1st Workshop on Network and System Support for Games, ACM, pp. 53–57 (2002)
4. Kim, S.-G., Lee, N.-J., Yang, S.-W.: A management method of load balancing among game servers based on distributed server system using map balance server. J. Adv. Navig. Technol. **15**(6), 1034–1041 (2011)
5. Takeshita, T., et al.: Introduction to Mastering TCP/IP, pp. 184–189. Ohmsha Ltd. (1998)
6. Fall, K.R., Richard Stevens, W.: TCP/IP Illustrated, Volume 1: The Protocols. Addison-Wesley (2011)
7. Wencong, W.J.X.C.W., Xiaofei, Z.: The application of IOCP in the P2P network game. Comput. Eng. Appl. **7**, 205–207 (2006)
8. Pantel, L., Wolf, L.C.: On the suitability of dead reckoning schemes for games. In: Proceedings of the 1st Workshop on Network and System Support for Games, ACM, pp. 79–84 (2002)
9. Casaló, L., Flavián, C., Guinalíu, M.: The role of perceived usability, reputation, satisfaction and consumer familiarity on the website loyalty formation process. Comput. Hum. Behav. **24**(2), 325–345 (2008)
10. Sandulli, F.D.: CD music purchase behavior of P2P users. Technovation **27**(6), 325–334 (2007)
11. Lee, K.C., Kwon, S.: Online shopping recommendation mechanism and its influence on consumer decisions and behaviors: a causal map approach. Expert Syst. Appl. **35**(4), 1567–1574 (2008)
12. Claypool, M., Claypool, K.: Latency and player actions in online games. Commun. ACM **49**(11), 40–45 (2006)
13. Cronin, E., Filstrup, B., Kurc, A.: A Distributed Multiplayer Game Server System. University of Michigan (2001)

Analysis of Large-Scale Diabetic Retinopathy Datasets Using Texture and Blood Vessel Features

Devvi Sarwinda, Ari Wibisono, Hanifa Arrumaisha, Zaki Raihan, Rosa N. Rizky FT, Rico Putra Pradana, Mohammad Aulia Hafidh and Petrus Mursanto

Abstract Diabetic retinopathy is a disease caused by the complications of diabetes mellitus and can cause blindness. In this study, we classified the stages of diabetic retinopathy using a large-scale dataset that consists of 35,126 fundus images. The classification of diabetic retinopathy includes five stages, from normal to proliferative diabetic retinopathy. In the proposed approach, a morphological feature extraction method and advanced local binary patterns were employed to extract blood vessel and texture features, respectively. The support vector machine, K-nearest neighbor, random forest, and logistic regression techniques were compared as classifiers. The classification was conducted on fundus images from a Kaggle dataset. The experimental results show that the texture feature extraction method based on advanced local binary patterns leads to higher accuracy, precision, and recall score than the blood vessel features extracted using morphological operators.

Keywords Diabetic retinopathy · Fundus images · Blood vessel · Texture feature · Large scale dataset

D. Sarwinda (✉)
Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Indonesia, Depok 16424, Indonesia
e-mail: devvi@sci.ui.ac.id

A. Wibisono · P. Mursanto
Faculty of Computer Science,
Universitas Indonesia, Depok 16424, Indonesia
e-mail: ari.w@cs.ui.ac.id

P. Mursanto
e-mail: santo@cs.ui.ac.id

H. Arrumaisha · Z. Raihan · R. N. Rizky FT · R. P. Pradana · M. A. Hafidh
Undergraduate Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia
e-mail: hanifa.arrumaisha@ui.ac.id

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_10

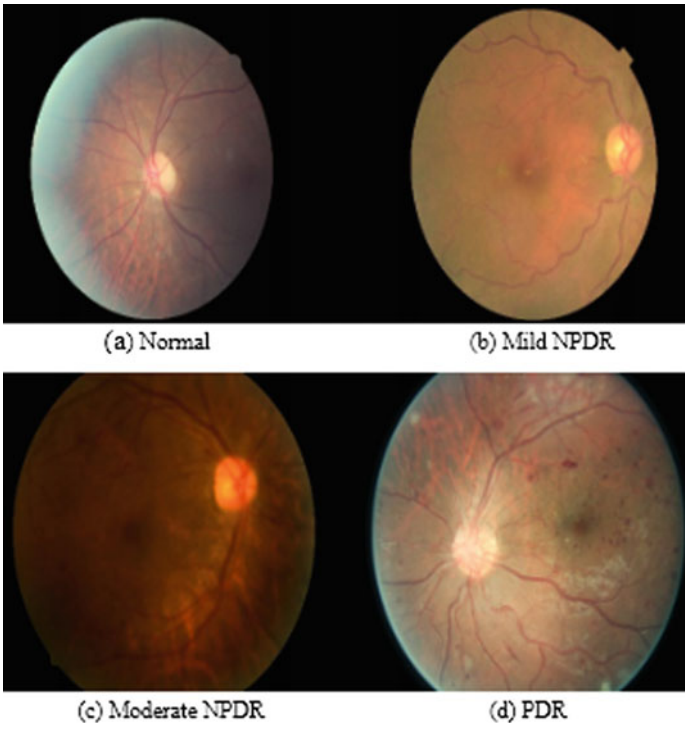


Fig. 1 Fundus images of the stages of diabetic retinopathy [8]

1 Introduction

Diabetic retinopathy is one of the main causes of blindness in people who have diabetes. According to International Diabetes Federation (IDF) data, there are around 425 million people in the world between the ages of 20 and 79 who have diabetes in 2017 [1]. This number is expected to increase to 48% in 2045. Based on the IDF data, 40−45% of diabetics also have diabetic retinopathy. Diabetes can create some complications in the nerve and blood vessels. Based on the level of development, there are two stages of diabetic retinopathy. The first or early stage is non-proliferative diabetic retinopathy (NPDR). NPDR consist of mild, moderate, and severe conditions. Pandelaki stated that NPDR is the initial stage of diabetic retinopathy where there are no apparent symptoms [2]. The next stage is proliferative diabetic retinopathy (PDR). Sufferers often experience ambiguous symptoms until the PDR stage is reached. This stage is characterized by the appearance of neovascular blood vessels. The stages of diabetic retinopathy using fundus images is shown in Fig. 1.

Many studies have been performed on the detection of diabetic retinopathy. Some use an image processing approach. For instance, Sarwinda et al. [3] proposed a method of detecting diabetic retinopathy by analyzing texture features using com-

plete local binary patterns and classifying them using the K-nearest neighbor (KNN) method. Akram et al. performed the detection of individual PDR stages using two-dimensional Gabor wavelets and multilayer thresholding, improving the accuracy of the previous method [4]. Verma et al. [5] detected each NPDR stage using density analysis and a bounding technique.

Several researchers detect each stage of the disease. For instance, Yun et al. classified diabetic retinopathy stages into normal, moderate NPDR, severe NPDR, and PDR [6]. They used the area and perimeter of each RGB component as a feature. Histogram equalization (HE) and binarization have been used to extract features and classify them using a neural network. The system achieved an accuracy of more than 80%, a sensitivity of 91.7%, and a specificity of 100%. Nayak et al. classified normal fundus images into NPDR and PDR stages. Three image features were used: blood vessels, exudates, and texture [7]. The three features were represented using adaptive HE (AHE) and classified using a neural network. Their method had a 71.4% specificity and 96.7% sensitivity.

Computer vision should be able to recognize or detect diabetic retinopathy stages simply and quickly. However, the detection of diabetic retinopathy stages is a challenging problem because the subsequent stages have almost the same lesions, although there are factors that distinguish them in theory. For this reason, it is necessary to choose a suitable method to classify diabetic retinopathy stages correctly.

In previous work [3–7], the researchers only used small-scale images. In this paper, we evaluate texture and blood vessel-based features on a large-scale diabetic retinopathy dataset. This study reports the accuracy, precision, recall score, and performance of the proposed method, especially for large data. The goal of this study is to classify diabetic retinopathy stages and gain knowledge about the characteristics of the progression of the disease. We used blood vessel and texture features as descriptors in our model.

An explanation of the material and methods used in this study (e.g., the data, preprocessing method, feature extraction method, feature selection method, and classification method) is given in Sect. 2. In Sect. 3, we describe the results of an evaluation of our proposed method and conclude our paper in Sect. 4.

2 Materials and Methods

2.1 Fundus Image Acquisition

Diabetic retinopathy fundus images were acquired from a Kaggle dataset [8]. The data consist of 35,126 images with five class labels (normal, mild NPDR, moderate NPDR, severe NPDR, and PDR). The Kaggle dataset has large proportion of uninterpretable images, which is due to faulty labeling and poor image quality. Hence, we evaluated several methods to determine the performance of each approach to the classification of diabetic retinopathy stages. The description and number of images for each label is shown in Table 1.

Table 1 Description of the five classes

Stage of diabetic retinopathy	Description	Number of images
Normal	Healthy retina	25,810
Mild NPDR	More than one sign of venous dilatation, microaneurysms, bleeding intraretina, and hard exudates	2,443
Moderate NPDR	More than one sign of venous dilatation bleeding, hard exudates, and soft exudates	5,292
Severe NPDR	Bleeding, microaneurysms in the fourth quadrant, and dilated veins in the second quadrant	873
PDR	Signs of NPDR along with the appearance of new or neovascular blood vessels	708

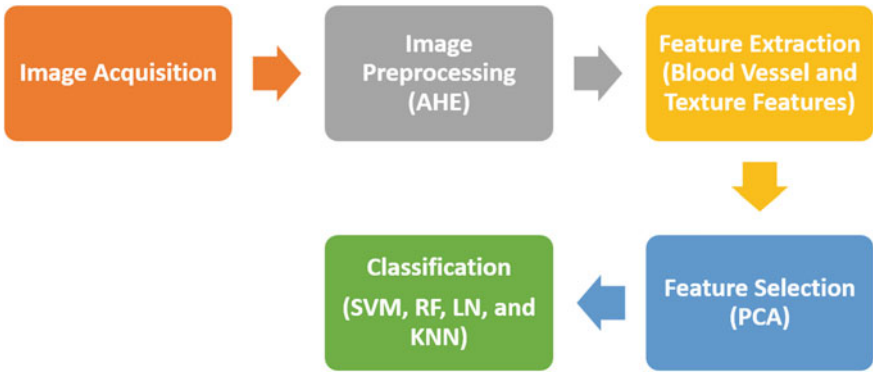


Fig. 2 Diabetic retinopathy stage classification system

2.2 *Research Methods*

The classification method used in this research is shown in Fig. 2. As the figure shows, our detection method extracts two types of features: blood vessel and texture features.

2.3 *Image Preprocessing Using the Green Channel and AHE*

In this step, the green channel of the RGB image is used because this channel produces a more stable image than the red and blue channels [9]. In this case, stable means that

the image produced is not too bright (it has moderate intensity levels) and not too dark (the contrast is also moderate) [9]. The green channel is also clearer so that it is suitable for medical image analysis. After the green channel is extracted, we invert it. The aim of this transformation is to best exhibit the vessels without the need to perform any further preprocessing.

The medication errors associated with diabetic retinopathy can limit the contrast of the lesion image and affect the detection process [10]. In these cases, one technique that can be used for histogram enhancement is AHE.

There are two approaches to modifying an image histogram. One of them is histogram leveling or HE, which changes the image intensity values to a uniform distribution. Histogram alignment is obtained by changing the gray value of a pixel (r) to a new gray value (s) using a transformation function T, that is, s = T (r). However, this technique is not sufficient to solve the problem, so AHE was proposed. The concept of AHE is almost the same as that of HE, but AHE performs an adaptive calculation per pixel to determine the new gray value. Many experiments have proven that AHE is useful for increasing image contrast [7].

In this study, AHE is used to produce a better image, and all areas of the image, both small and large areas, are included in the contrast uniformity. Then, the preprocessed image is resized to 845 833 pixels. The stages of image preprocessing are shown in Fig. 3. While, the stage of image processing for each classes are shown in Figs. 4, 5, 6, and 7.

2.4 Feature Extraction Using Morphology

Generally, morphology expresses form. In the context of mathematics, morphology is a feature extraction method that is useful for the representation and description of sectional shapes, such as boundaries, skeletons, and convex hulls. This technique is also used for preprocessing and postprocessing. The basic morphological operations are dilation and erosion, which can be used to create opening, closing, top-hat, and bottom-hat transforms. In general, morphological operations pass a structuring element over an image and are used to examine the original image. A structuring element must have a center point. Some examples of structuring element types are arbitrary, diamond, disk, line, octagon, pair, periodic line, rectangle, and square shapes.

Figure 8 illustrates the morphological extraction of blood vessel features. In this process, there are five steps: opening, optical disk removal, thresholding, median filtering, and area calculation. In the first step, we performed an opening because the results we obtain are better than those of other morphological methods. After opening, an image appears more arising and has less noise. An opening is an erosion followed by a dilation. Based on the results of preliminary experiments, we used a 17×17 ball for the structuring element.

Next, a dilation was performed, which joins separate image components (using element subtraction) into a whole component. In the dilation process, we subtracted

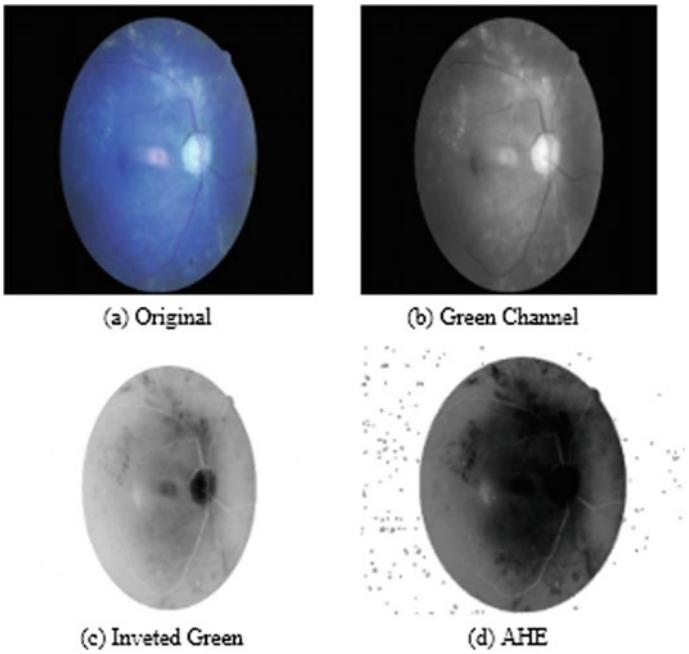


Fig. 3 Stages of image preprocessing

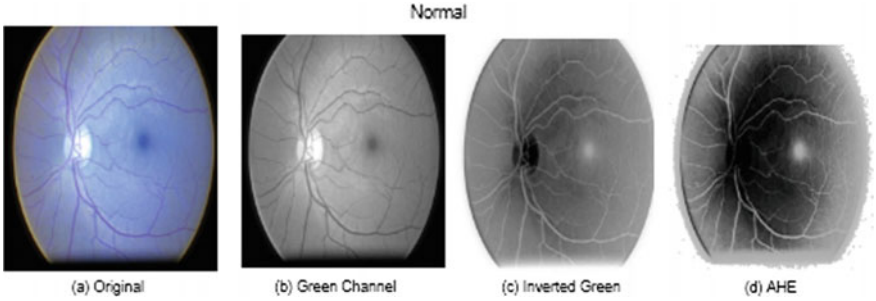


Fig. 4 Stage of image preprocessing for normal dataset

the opened image to remove the optical disk. This is because the optical disk is the part of the retina where the optical nerve forms and does not indicate diabetic retinopathy in the retina. Then, we performed a thresholding of the image after subtraction to obtain its area features. This process aims to clarify the blood vessel curves in the image. The threshold value was set to 50 because this value was considered to be the best in our experiments. Because the results still contain many dots that are considered noise, they were removed using a median filter.

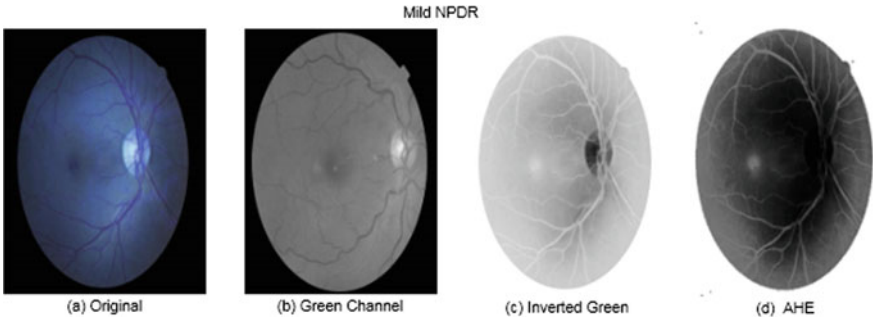


Fig. 5 Stage of image preprocessing for mild NPDR dataset

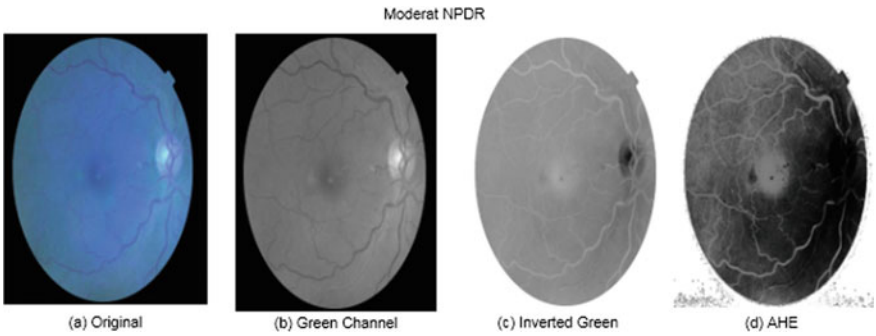


Fig. 6 Stage of image preprocessing for moderate NPDR dataset

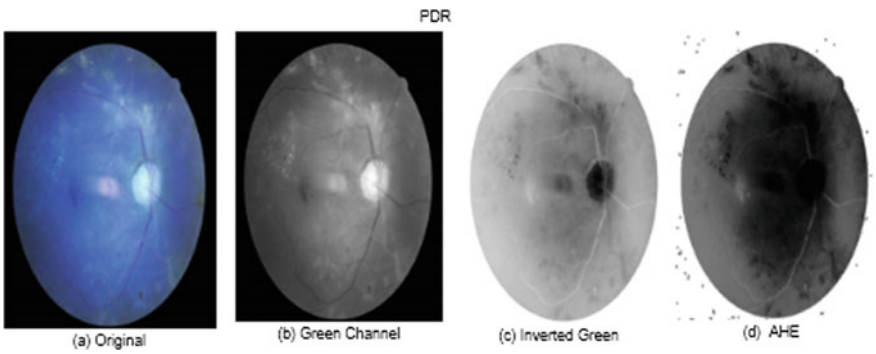


Fig. 7 Stage of image preprocessing for PDR dataset

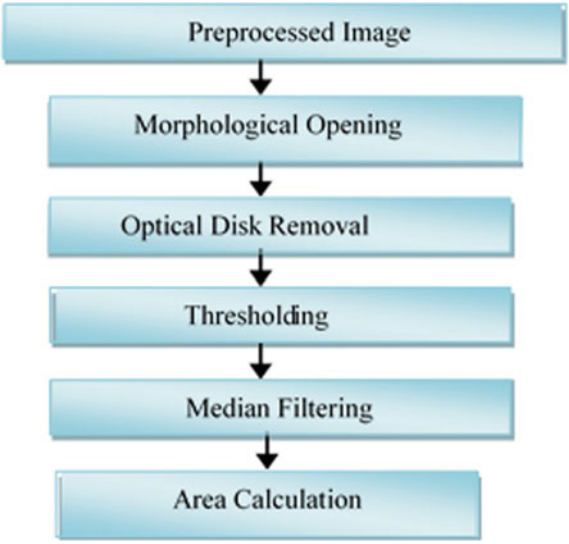


Fig. 8 Feature extraction process for the blood vessel feature

2.5 Feature Extraction Using Advanced Local Binary Patterns (ALBPs)

The advanced local binary pattern (ALBP) method is an extension of the local binary pattern (LBP) method [11], which analyzes the texture of an image based solely on a conversion of the gray intensity values to binary values as a feature. The ALBP method uses the binary values from the gray intensity values as a feature, but then adds the values obtained by the magnitude value (called the LBP magnitude). The result of combining the binary and magnitude values produces a new feature called the ALBP feature [12].

The magnitude (m_i) of the ALBP is the absolute value of the difference in the intensity of the gray value (δp_i). Let the average (c) of the magnitude value be the center. Based on the value of neighbor m_i and center c, the LBP magnitude is calculated, where the binary value is M_i. The calculation is mathematically shown as follows.

$$LBP_{x,y}magnitude = \sum_{i=1}^{n} 2^i . M_i(m_i - c) \tag{1}$$

where

$$M_i(m_i - c) = \begin{cases} 1, \text{ if } m_i \geq c \\ 0, \text{ if } m_i < c \end{cases} \tag{2}$$

and

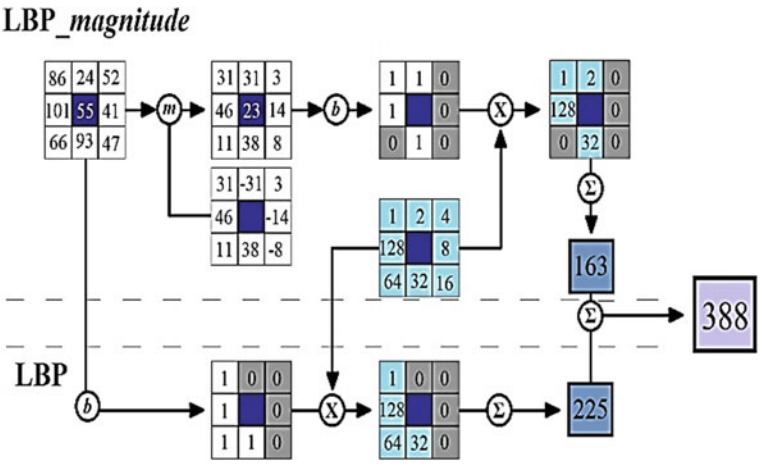


Fig. 9 Example of ALBP

$$\left(M_i|m_i = |\Delta p_i|, \quad c = \frac{1}{n}\sum_{i=1}^{n} m_i\right)$$

Hence, the ALBP value can be obtained by summing the value of an ordinary LBP and the LBP magnitude as follows.

$$ALBP_{x,y} = \sum_{i=1}^{n} 2^i.(M_i(m_i - c) + B_i(\Delta p_i)) \tag{3}$$

where

$$B_i(\Delta p_i) = \begin{cases} 1, \text{ if } \Delta p_i \geq 0 \\ 0, \text{ if } \Delta p_i < 0 \end{cases} \tag{4}$$

$$\Delta p_i = p_i - p_{\text{pusat}(x,y)} \tag{5}$$

An illustration of ALBP can be seen in Fig. 9. In this study, we used an ALBP with rotation uniform invariance to extract texture features from fundus images. In previous work, ALBP with rotation uniform invariance outperformed other approaches of LBP [12]. Our method uses the preprocessed AHE images as the input for feature extraction.

2.6 Feature Selection Using Principal Component Analysis (PCA)

Principal component analysis (PCA) is a method for building linear combinations of original variables into new variables, which are called principal components [13]. These new variables are not correlated and arranged so that the first principal component is the component that has the highest variance. The steps performed in the PCA are as follows:

1. Suppose data X can be formed in a matrix as follows:

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nn} \end{bmatrix}$$

   where n is the number of variables and m is the number of observations.
2. Then, the data is transformed into a column by reducing each data by the average of each attribute as follows.

$$\hat{X} = X - \bar{X} \tag{6}$$

   Here, $\hat{X}$ is the resultant vector after centering, X is the column vector, and $\bar{X}$ is the average of the column vector.
3. The covariance for Eq. 6 is calculated.
4. The eigenvalues and eigenvectors are then determined. The principal components are obtained from the eigenvector.

2.7 Classification

In this study, the classifiers are used to classify classes are a Support Vector Machine (SVM), KNN, Logistic Regression, and a Random Forest. For the classification process, we set different hyperparameters for each classifier (SVM used the Radial Basis Function Kernel (RBF), KNN used K = 5, and the Random Forest had a tree size of 100). To evaluate our models in both approaches, we measured model performance with respect to accuracy, precision, and recall.

The formulas for accuracy, precision, and recall are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F - 1 = \frac{2(Precision \times Recall)}{Precision + Recall} \tag{10}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

In the medical context, recall has the same definition as sensitivity, where it is a basic method for determining the proportion of true positive cases that are correctly detected as positive [14]. In contrast, precision is defined as the proportion of predicted positive cases that are real positives [15]. F−1 score is one of the derivatives of the F-measure which describes harmonic mean of precision and recall. 1 in F−1 score stands as a parameter that balance the precision and recall scores. Formula 10 given below defines the F−1 score with P stands for the precision score and R stands for recall score.

3 Results and Discussion

In this section, we describe the results of our experiment, which consisted of two scenarios. In the first scenario, we classified our data using the blood vessel feature approach, and in the second scenario, we used the texture feature approach to classify the diabetic retinopathy stages.

In the experimental process, we performed image preprocessing, feature extraction, feature selection, and classification using the various methods.

For the feature extraction, we extracted blood vessels from the image, as shown in Fig. 10. Next, we performed feature selection using PCA to reduce the dimensions of the image features. Because the blood vessel extraction produced high-dimensional features, before the PCA process started, we changed the image resolution 100 × 100 pixels to avoid out-of-memory errors on our devices. In blood vessel extraction, we select features from PCA with various number of features, such as 50 features, 10% of features number, 20% of features number, and 30% of features number. From four kinds of features number are selected, we obtained the best evaluation value (accuracy, precision, recall, and time) from 50 features. Comparison of running time for each of classification in various features are described in Table 2. From Table 2, feature selection using PCA for 50 features has lower time than another features for whole classifiers. We also performed the same steps using the ALBP method to extract texture features. Using ALBPs, we obtained 100 features, and we selected 20 of these features using PCA.

Both types of features were classified using the same classification methods (SVM, KNN, logistic regression, and random forests). Based on the following scheme, the classification was carried out. For both feature extraction methods, we split the data 75:25 for learning such that 75% was for training and 25% was for testing. The results

Table 2 Comparison of running time for various features by PCA in blood vessel features

Model	PCA 50	PCA 10%	PCA 20%	PCA 30%
SVM	309 s	4031 s	5517 s	5335 s
RF	26 s	116 s	160 s	197 s
LR	1 s	39 s	103 s	185 s
KNN	18 s	352 s	588 s	861 s

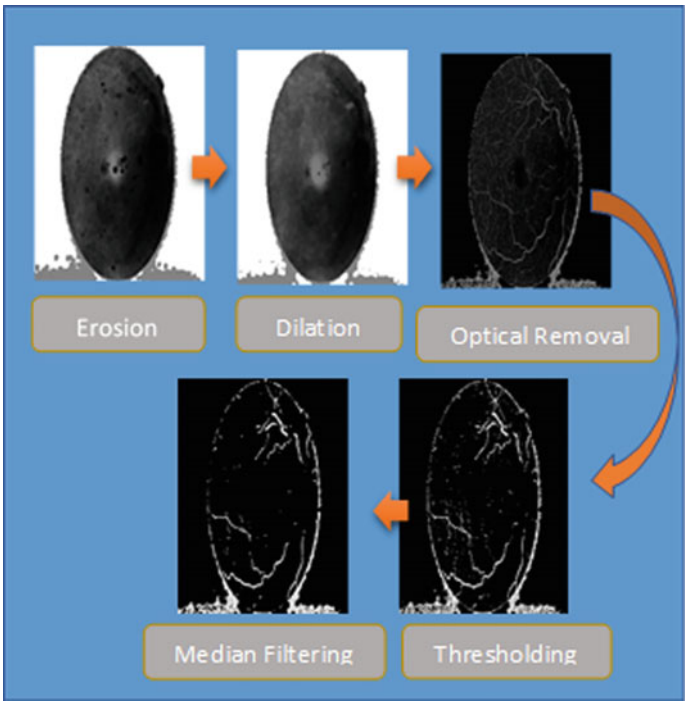


Fig. 10 Example of blood vessel feature extraction for a moderate NPDR fundus image

of the accuracy, precision, and recall are compared in Figs. 11 and 12. In Fig. 11, four classifiers show almost same values for classification accuracy. When using blood vessel extraction, SVM performs well with respect to other classifiers overall (73.49, 34.7, and 20% for accuracy, precision, and recall, respectively). Meanwhile, Fig. 12 shows that higher accuracy, precision, and recall were obtained by the combination of texture features and the random forest classifier (82.27, 83, and 77%, respectively). In additional evaluation, we also evaluate our proposed method using F−1 score. F−1 score describes harmonic mean of precision and recall. Figure 13 shows that texture features has higher F−1 score than blood vessel extraction.

The extraction times for both approaches were measured. The data processing stage was performed on a Triple Nvidia GTX 1080 32 GB server with Intel® Core™

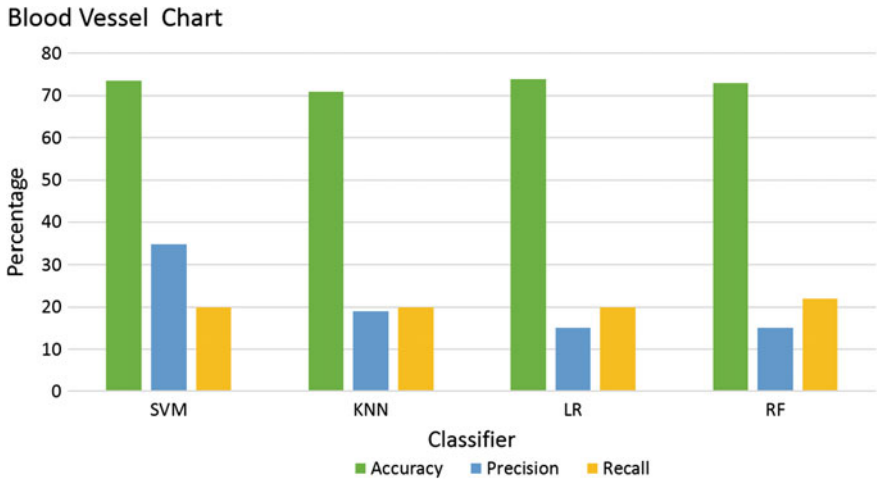


Fig. 11 Performance results of the classifiers using blood vessel features

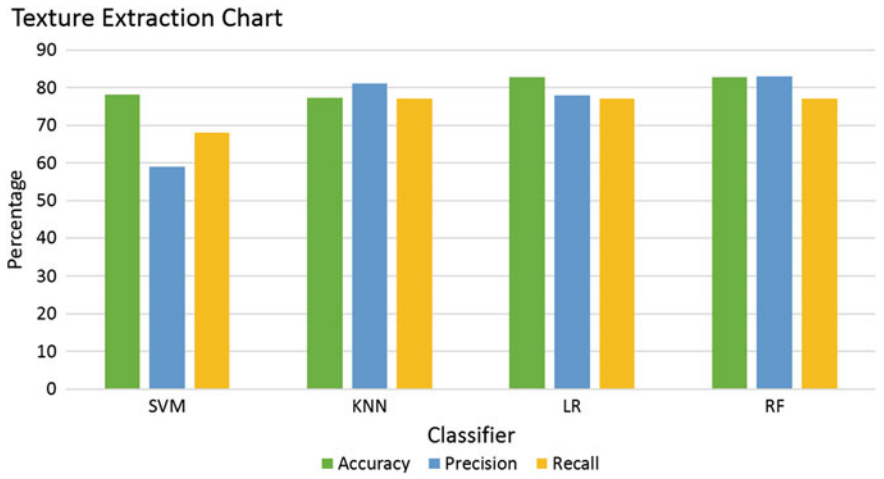


Fig. 12 Performance results of the classifiers using texture features

faster than the extraction of the ALBPs for the texture feature. The morphological processing needed 4,500 s for blood vessel extraction, and the ALBP method took 36,000 s.

The results of the blood vessel and texture features approaches indicate that texture is also an important feature for detecting diabetic retinopathy, although its extraction time is higher than that of blood vessels. This is because the texture feature was processed for the whole image without segmenting out component images as in blood vessel extraction. These results further indicate that this approach could be used as a recommendation system to diagnose diabetic retinopathy.

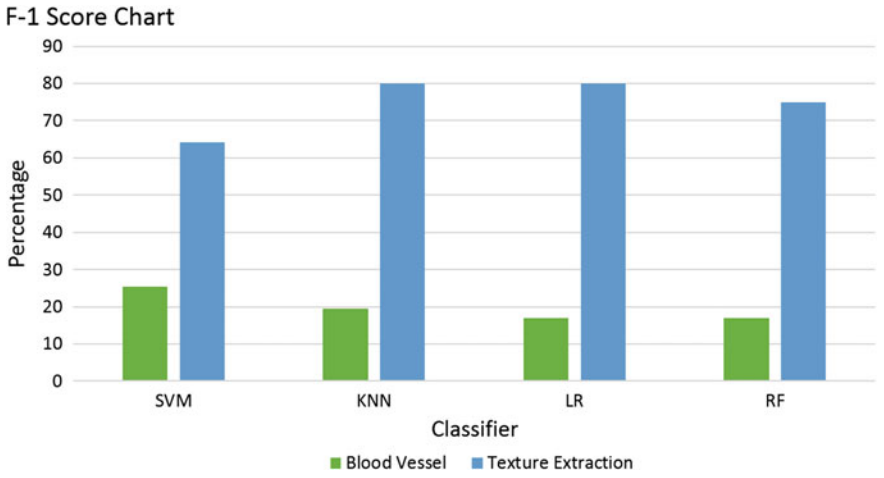


Fig. 13 F−1 Score chart

4 Conclusion and Future Work

Diabetic retinopathy is a disease caused by diabetes complications and is a major cause of blindness. Early detection is useful for reducing the risk of blindness. In this study, different feature extraction methods were evaluated for the detection of diabetic retinopathy stages. Both approaches can classify the stages of diabetic retinopathy. Our experimental results for the accuracy of the four-class classification are sufficient to distinguish one class from the other classes in a large-scale dataset. This is caused by imbalanced data, different detection factors for each class, and the misscategorization of lesion characteristics. Overall, for future research, we suggest further experiments with other feature extraction methods to analyze big data images and other medical images. We could also improve the performance using advanced computing [16].

Acknowledgements We would like to express our gratitude for the PIT 9 2019 Grant from Directorate of Research and Human Engagement Universitas Indonesia with contract number is NKB-0011/UN2.R3.1/HKP.05.00/2019 in supporting this research. Moreover, we would like to thank Kaggle for providing the dataset.

References

1. International Diabetes Federation.: IDF Diabetes Atlas, 8*th* edn. (2017)
2. Pandelaki, K.: Buku Ajar Ilmu Penyakit Dalam Edisi 4. Jakarta, Departemen Ilmu Penyakit Dalam FKUI (2007)
3. Sarwinda, D., Bustamam, A., Wibisono, A.: A complete modelling of local binary pattern for detection of diabetic retinopathy. In: Proceedings of the International Conference on Informatics and Computational Sciences (ICICoS), pp. 7–10. IEEE (2017)

4. Akram, M., Usman, et al.: Automated segmentation of blood vessels for detection of proliferative diabetic retinopathy. In: Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), vol. 2–7, pp. 232–235. IEEE, Hong Kong and Shenzhen, China (2012)
5. Verma, K., Deep, P., Ramakrishnan, A.G.: Detection and Classification of Diabetic Retinopathy Using Retinal Images. In: Annual IEEE India Conference (INDICON), pp. 1–6. IEEE (2011)
6. Yun, W.L., et al.: Identification of different stages of diabetic retinopathy using retinal optical images. Inform. Sci. **178**, 106–121 (2008)
7. Nayak, J., et al.: Automated identification of diabetic retinopathy stages using digital fundus images. Springer J. Med. Syst. **32** (2), 107–115 (2008)
8. Diabetic Retinopathy Dataset. https://www.kaggle.com/c/diabetic-retinopathy-detection/data
9. Fitriati, D., Murtako, A.: Implementation of diabetic retinopathy screening using real time data. In: Proceedings of the IEEE 2016 International Conference on Informatics and Computing (ICIC), pp. 198–203 (2016)
10. Datta, N.S., Dutta, H.S., De, M., Mondal, S.: An effective approach: image quality enhancement for microaneurysms detection of non-dilated retinal fundus image. Procedia Technol. **10**, 731–737 (2013)
11. Pietikinen, F.M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns. Springer, London, pp. 13–27 (2011)
12. Sarwinda, D., Bustamam, A.: Detection of Alzheimer's disease using advanced local binary pattern from hippocampus and whole brain of mri images. In: International Joint Conference on Neural Networks (IJCNN), pp. 5051–5056 (2016)
13. Platt, J.C.: Fast Training of Support Vector Machines Using Sequential Minimal Optimization. Advances in Kernel Methods, vol. 12, MIT Press Cambridge, pp. 185–208 (1999)
14. Powers, D.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. J. Mach. Learn. Technol. **2**, 37–63 (2007)
15. Parikh, R., et. al.: Understanding and using sensitivity, specificity, and predictive value. Indian J. Ophthalmol. **56**(1), 45–50 (2008)
16. Bustamam, A., et al.: A GPU Implementation of fast parallel markov clustering in bioinformatics using ELLPACK-R sparse data format. In: International Conference on Advances in Computing Control and Telecommunication Technologies (2010)

Combination of 1D CNN and 2D CNN to Evaluate the Attractiveness of Display Image Advertisement and CTR Prediction

Wee Lorn Jhinn, Poo Kuan Hoong and Hiang-Kwang Chua

Abstract With the explosion of digital data nowadays, it has catapulted the usage of data analytic in the emergence of digital advertising space. One of the digital advertising giants, Facebook has accelerated the growth of this digital data volume as they are the most common used platforms for advertisers to advertise and deliver advertising messages to the mass audience online. However, this phenomenon has increased the challenges faced by advertisers to further attract audiences attentions to look at the digital advertisements when they are shown advertisements in Facebook platforms. Hence, in this paper, we proposed a method to evaluate and analyze the elements of attractiveness within the display advertisement in Facebook Advertisement platform by applying the 2D CNN on the display advertisement images while 1D CNN on the click metric data respectively. Based on our experiment results, we are able to predict the display CTR with a reasonable margin of error.

1 Introduction

In recent years, digital advertising has turned into a multi-billion dollar industry and begun to show tremendous impact in terms of generating a notable amount of revenue for ones company or raising general awareness to the public. Generally, there are mainly two types of digital advertising methods, namely: display advertising and search advertising. It is often a struggle for advertisers, particularly, to decide the placement type of the advertisement, in order to serve a better fit to the campaign objectives. Search advertising is known to provide a better chance of approaching

W. L. Jhinn (✉) · P. K. Hoong · H.-K. Chua
Axiata Digital Advertising (ADA), Level 32, Axiata Tower, Jalan Stesen Sentral 5, Kuala Lumpur Sentral, 50470 Kuala Lumpur, Wilayah Persekutuan, Malaysia
e-mail: lornjhinn.wee@ada-asia.com

P. K. Hoong
e-mail: poo.kuanhoong@ada-asia.com

H.-K. Chua
e-mail: george.chua@ada-asia.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_11

a potential customer more easily and thus, creating an opportunity of penetration throughout the emulation with multiple competitors in the digital advertising platform. Nonetheless, it is worth noticing that the search advertising only retains this customer penetration advantage under a circumstance, where a customer knows well in what category it is aiming for. On the other hand, display advertising focuses more on driving brand awareness. Since display advertisement can target specific audience segments based on the predefined demographic, locations, and the keywords within the surfed internet content, it can help to attain first impression. It has been proven that the first opportunity to see contributes 73% of the short-term sales effect of advertising [1].

Furthermore, it is forecasted by Cisco that 80% of the contents in the internet are going to be occupied by image based contents and it is strongly believed that this will be growing exponentially until year 2022 [2]. Consequently, the aesthetic design and context in an image for the digital display advertisement begin to play an essential role in capturing ones attention. Besides a well execution on an advertisement campaign, a generally correct interpretation for the display advertisement message based on the visual context is crucial in order to stimulate a series of desired post action (e.g.: conversion and brand awareness). It is believed that the image can generate a stronger impact in online advertising [3]. Moreover, the rise of big data era in digital advertising has created a revamp for the Moores law effect, where there is a massive yet swift incrementation in digital advertising data volume, in which is eagerly demanding a more sophisticated architecture design on computation hardware, specifically Graphical Processing Unit (GPU) to optimize the computation time needed by different algorithms for evaluating the advertisement campaign performance. Subsequent to that, more prospective algorithms are needed to extract a deeper insight of the data while fully leveraging the GPU processing power.

Although Human Intelligence (HI) may unfold some undiscovered attractive elements in a display image occasionally and evaluate the effectiveness of the display advertisement, there are approximately millions of historical data that are being generated simultaneously in different regions each day. With the assistance of Deep Neural Network (DNN) technique, it helps to reduce the burden of HI in evaluating over a huge amount of the same category image data to verify a similar effectiveness on the similar display advertisement. However, the challenge faced is that how optimum can the DNN technique be "taught" in order to evaluate the effectiveness of the display advertisement by extracting the attractive elements in an image data, humanly.

Therefore, in this paper, we propose a DNN method by applying 2-Dimension (2D) Convolutional Neural Network (CNN) and 1-Dimension (1D) CNN on the Facebook display images click metrics dataset, respectively, to predict the Clickthrough Rate (CTR) as CTR has been one of the popular measures for evaluating the efficiency of online advertising [4]. It is noted that the proposed method is mainly focusing on optimizing the CTR regression value as the good CTR for different advertisement categories varied in the applied Facebook dataset. Thus, to avoid any induction of human biasness (predefined CTR threshold) towards a good or bad CTR, Mean Square Error (MSE) loss will be used in this study as the evaluation metric to evaluate the predicted CTR value derived by the proposed method. It is noted that a good or

bad CTR should not be defined based on whether the CTR is higher or lower than the predefined CTR threshold value in the dataset but it should follow the average CTR of different advertisement categories as a threshold respectively.

2 Related Works

There have been several proposed techniques that applied CNN to CTR prediction based on the click metrics data. In 2007, Richardson et al. showed the effectiveness of logistic regression model on predicting how likely is the search advertisement will be clicked [5]. In 2015, Liu et al. proposed a CNN based CTR prediction model, where it can handle a varied length types of input instances as each sample contains a derived one-hot encoded vector based on the click metric data [6]. Nonetheless, this leads to a high dimensionality and sparse input space for the CTR model [7]. In 2016, Qu et al. proposed a Product Neural Network (PNN) that is able to capture the high order feature interaction by including a product layer after the embedding features are generated [8]. Meanwhile, in 2018, Patrick et al. presented a random multiple sequence embedding feature vector combination to study the influences of the sequential characteristic in embedding feature towards CTR prediction [9]. Generally, these methods shared a similar model structure design, where the embedding layers are firstly adopted to derive a dense representation from sparse features and followed by attaching a Multilayer Perceptron (MLP) layer to learn the relationship between the different combinations of features [10]. Zhou et al. introduced a DIN (Deep Interest Network) and an efficient mini-batch aware regularizer where DIN serves to learn the significant representation of user interests based on the historical behaviors w.r.t specific display while the mini-batch aware regularizer technique applies the L2-norm regularization on parameters that generate non-zero features in order to speed up the training process in industry level deep networks [10].

As for the studies on display images, Cheng et al. analyzed and reported a positive impact on CTR prediction accuracy by integrating the multimedia features from display images into the historical click metrics data [11]. Subsequently, Chen et al. in 2016 proposed a DeepCTR, which is the combination of ConvNet (17 layers proposed CNN architecture model) on display image and BasicNet (one-hot-encoding MLP) to reduce the dimensionality on one-hot encoded basic click metrics data. The CombNet is said to have the ability in learning complex yet effective non-linear from these two features [12]. In recent research, inspired by Chen et al., Michel et al. proposed an alternative method with a similar CNN architecture by applying ImageNet 1000 class log probabilities as CNN input to explore the influence on CTR classification [13].

3 Proposed Methods

In this section, the architecture design of the proposed 2D CNN and 1D CNN are illustrated and explained accordingly. Following that, we introduce several types of the global image feature extraction techniques that are adopted into this study. Based

on our experimental results, the extracted global image features from the display images do help in improving the proposed CNN model CTR prediction value.

3.1 2D CNN Network Architecture

The design of the 2D CNN is mainly inspired by the neural network architecture in [12, 13]. With the similar view as [12] that the trade off in between the training duration and performance must be considered. It is noted that with the consideration of overall performance, we did not build a very deep network architecture. Besides, we found that the fine-tuned transfer learning model such as ResNet for display advertising [13] tends to have a limitation on extracting the useful high-level image features from the cartoonized display images in our Facebook dataset. This is due to the fact that most of the state-of-the-art transfer learning models were developed mainly to classify only the real-life objects. Therefore, it is imperative to build a customized 2D CNN network, which is able to help in extracting not only the real-life objects, but also partially the "cartoonized" display image features that would be able to ameliorate the learning in predicting the CTR value. Figure 1 illustrates the designed architecture.

In this network, it consists of 13 convolution layers. The first convolution layer is a 5×5 convolution kernels in order to generate a higher-level feature map on the first hand. After that, there are three blocks of convolution operations, where each block contains of four convolution layers with a 3×3 convolution kernels individually. Subsequently, each convolution block operation, a 2D max-pooling is performed to down-sample the input image by summing up the activation matrix based on two hyperparameters known as: filter and stride, where the filter, F is set to have a standard value of 2 in order to sum up the feature values in each 2×2 matrix region without overly destructing the potential useful image features while stride, S with a standard value of 2 to down sample every depth slice of the activation output matrix. The updated size, $W \times H \times D$, for the image activation output matrix can be defined as:

$$W_i = \frac{W_{i-1} - F}{S} + 1 \tag{1}$$

$$H_i = \frac{H_{i-1} - F}{S} + 1 \tag{2}$$

where D remains to be the same three image dimension.

3.2 1D CNN Network Architecture

Unlike most of the state-of-the-art techniques that convert the basic click metrics data into an embedding features vector. In this paper, the proposed 1D CNN fully leverages the numeric correlation between these non-linear click metrics to extract the linear

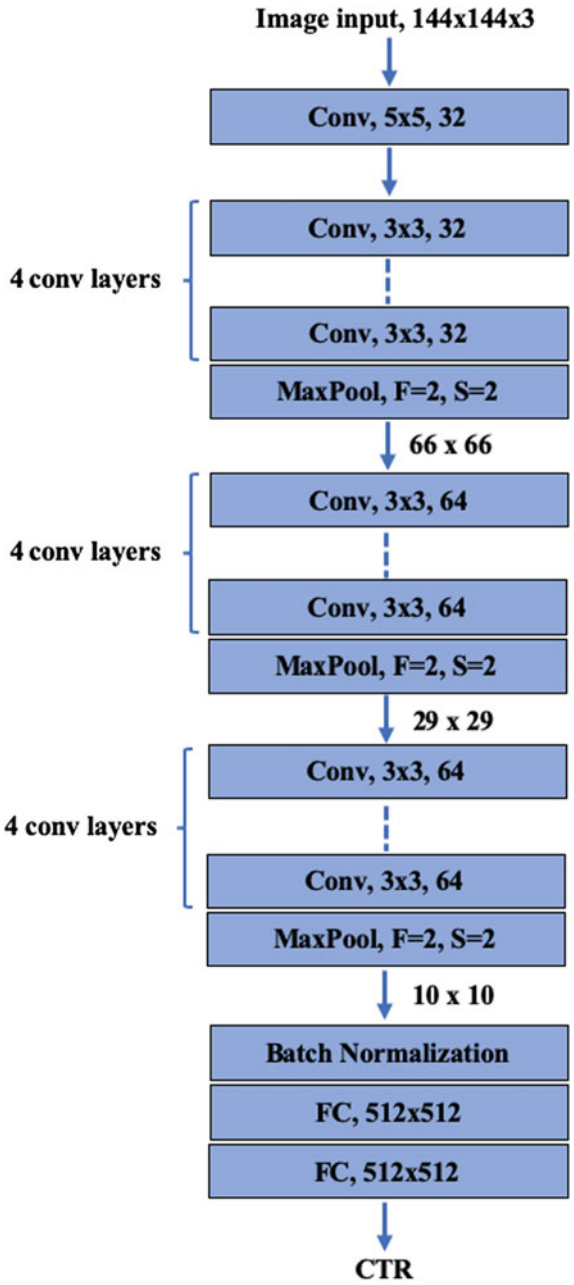


Fig. 1 The proposed architecture of the 13 convolution layers

relationship. Moreover, the striding and filtering properties of the CNN architecture contribute to a vast reduction in training parameters as compared to the general linear regression model. Consequently, the model convergence speed can be greatly accelerated, which is essential to perform a fast iteration on parameters fine-tuning, especially on a large volume dataset. It is worth mentioning that the Facebook dataset used in this study has a significant different sequential characteristic compared to the datasets in the other related work techniques. In [6, 9, 10, 12], they have the advantage of first party data, where the time sequential relationship between click metrics based on each customers historical data is apparent. However, for our Facebook dataset, which served as a second party data, contained only the accumulated value for each corresponding click metrics data, hence losing the valuable insight on sequential click metrics information of each customer. Therefore, 1D CNN can be solely acts as an alternative to predict the CTR based on the large-scale second party dataset.

Given a list of Facebook dataset input instance with n elements (metrics), denoted as $S_{i_{n=1,2,3...74}} \varepsilon R^{d \times i}$, the input instance matrix, x can be formed as below:

$$x = \begin{bmatrix} \vdots & \vdots & \vdots \\ S_i & \dots & S_n \\ \vdots & \vdots & \vdots \end{bmatrix} \tag{3}$$

Note that the element values in the x have been pre-processed by applying a min-max normalization to produce a normalized matrix, z, in order to reduce the sparse variations in between different input instances:

$$Z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4}$$

After that, the convolution process begins by convolving with the weight matrix, $w \varepsilon R^{i \times n}$, to generate the final activation output matrix, r in a one-dimensional way. To be specific, given the receptive field, $w_f \varepsilon R^{i \times n}$ of height in between $w_f = \{32, 64, 128\}$ instead of i and the number of kernel for each convolution, $k_{l=1,2,3}$ in between $2 \leq k \leq 10$, where l represents the lth designed convolutional layer, the lth layer's activation matrix output, z_l is computed by applying the non-linear activation function known as Rectified Linear Unit (ReLU) during the convolutional operation (Fig. 2):

$$Z_l = ReLU(w_f \cdot z_{l-1}) \tag{5}$$

where ReLU function only accepts the positive z_i as the features while denoting the negative or zero value found in z_i as 0 during the convolution operation. After each convolution operation, a 1D max-pooling will be performed to down-sample the input instance. The updated output volume, V for each layer of z_l can be defined through the calculation in below:

Total Facebook Data Records	130, 000
Total Unique Image	2,557
Facebook metric data (Total 74 types of features)	account_currency, impressions, clicks, post_reach, reach, spend, page_engagement, frequency, ctr, link_click, comment etc.
Image global features	brightness, saturation, colorfulness, naturalness, grayscale
Brightness	Average: 136.611 Standard Deviation: 61.179 Max: 255 Min: 3.262
Saturation	Average: 0.392 Standard Deviation: 0.249 Max: 1 Min: 0
Colorfulness	Average: 136.611 Standard Deviation: 61.179 Max: 255 Min: 3.262
Naturalness	Average: 0.502
Grayscale	Average: 62.811

Fig. 2 Summary of the applied Facebook dataset

$$V_{update} = \left(\frac{n - filter}{stride} + 1\right) \times w_f \tag{6}$$

The overall architecture of 1D CNN is illustrated in Fig. 3.

3.3 Image Global Features

Despite of the combination in between the 2D CNN and 1D CNN to improve the learning process between the complex non-linear features, some global image feature calculations are adopted to enhance the CTR prediction by establishing a direct correlation between the image features and click features. The applied global features are listed in the following:

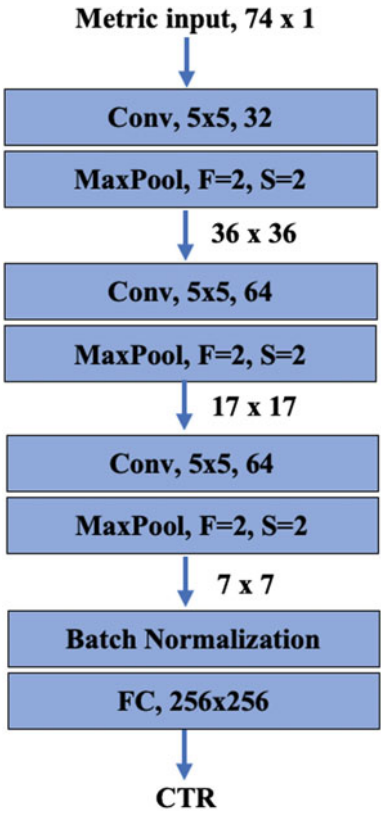


Fig. 3 The proposed architecture of the 3 convolution layers

- ***Brightness***: The image brightness value can be obtained directly from two color spaces known as: YUV and HSL, where the brightness in YUV is the luma component, denoted as Y channel among the YUV and the lightness, denoted as L in HSL. The average, standard deviation, maximum, and minimum of the brightness values of each display image are computed and added into the click metrics data.
- ***Saturation***: The image saturation indicates the vividness of an image. However, the image saturation features can be extracted directly from the HSL or HSV color space models. The average, standard deviation, maximum and minimum of the saturation values are added into the click metrics data.
- ***Colourfulness***: The image colorfulness features, C measures the differences against gray scale colours through the computation in RGB color space [14].

$$C = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{7}$$

where,

$$rg = R - G \tag{8}$$

$$yb = \frac{R + G}{2} - B \tag{9}$$

- ***Naturalness***: Naturalness denoted as N, measures the correspondence degree in between the images and human visual perception. This technique is proposed by Huang et al. [15] by grouping the pixel values based on the threshold of $20 \leq L \leq 80$ and $S > 0.1$ in the HSL color space and followed by defining the pixel values into three categories namely: (1) Skin, (2) Grass, and (3) Sky as the quantitative description. The final naturalness score is defined as:

$$N = \sum_{i} NS_i \times NP_i,\ i\ \varepsilon\ \{Skin, Grass, Sky\} \tag{10}$$

  where NP indicates the proportional pixel of the corresponding categories.
- ***Grayscale***: The grayscale value, G of the image is computed in the RGB channel by:

$$G = 0.299R + 0.587G + 0.114B \tag{11}$$

The standard deviation of the grayscale features is computed and added into the click metrics data. Javad et al. showed that this grayscale level features are very effective in predicting the CTR of advertisement [16]. Thus, there is an additional of 14 image global features are added into the click metrics data.

4 Experiments

The proposed model is evaluated in this section. Firstly, the Facebook dataset used in this study is explained progressively. Follow by illustrating the details of the experiment setup for model training on Facebook image data and evaluation metrics. Lastly, the visualization on the image salient features based on the proposed 2D CNN and the comparison results of between the proposed 1D CNN with global image features and without global image features are depicted with illustrations. The overall summary of the Facebook dataset is shown in Fig. 2.

4.1 Experimental Setup

In this study, the Facebook dataset with the total of 2,557 unique display advertisement images from the 130,000 click metrics were used. For display advertising, it is a common practice to create multiple advertisement campaigns with the same group of images in order to serve the same objective such as: raising public awareness

to certain topic or triggering post conversion for instance: click on the button or hyperlink to drive the website visitation rate, purchase of item or sign up for the campaign event. As mentioned before, the characteristic of the Facebook dataset is different compared to the other datasets in terms of the sequential data where a total of 130,000 metric data were taken for the period of 6 months ranging from August 2018 to January 2019. The selected Facebook dataset varies in regards to CTR based on the different click metrics data but with the same corresponding single display image.

Nevertheless, a low CTR display advertisement does not fully imply low attractiveness level of the display image as it is partially affected by the period (e.g.: weekly or monthly) of the advertisement campaign. Therefore, it will lead to a different CTR value for different campaigns, even though these campaigns served under the same objective. To address the problem of different CTR of the same image during the 2D CNN training, the varied CTR values on a single display image are averaged before passed into the 2D CNN model. On the other hand, the 130,000 metrics data with global features are passed instantly into the 1D CNN model without averaging the CTR as each metrics data has a different features value combination leading to different CTR compared to image features.

In this experiment, the Facebook dataset is divided into 90% for training and 10% for testing respectively. During the training, the Adam optimizer is adopted with a learning rate of 0.01 and batch size of 32 to prevent the overfitting in both CNN models. The predicted CTR value is evaluated based on MSE. A NVIDIA Geforce GTX 1050 Ti with Max-Q Design GPU processor that contains a 4GB memory is used to conduct these experiments.

4.2 2D CNN Experimental Results

From Fig. 4, it illustrates the saliency maps of three test display images after the proposed 2D CNN is applied. Throughout the experiment, it is showed that the saliency map of the pre-trained 2D CNN with the image size of 144 (second column), denoted as 144-saliency map, displays more robust salient features compared to the saliency map with the image size of 192 (third column), denoted as 192-saliency map that induced the image features as noise inadvertently. For instance, it can be observed that the 144-saliency map on the first image visualizes the little boys head at a superficial level while visualizing more at the table region. Meanwhile, the 192-saliency map overly visualized the region at the background yellow color wall. As for the second image, the 144-saliency map protrudes the 8 GB text and the person while the 192-saliency map specifies in terms of visualizing the “8” from the “8G”. Although the saliency maps of both image size in the third image seemed to be displayed at the similar region, it is observed that the visualization on the cartoon character in 144-saliency map tends to show a thicker focus region compared to 192-saliency map. Therefore, through this empirical test, we have concluded to train the 2D CNN with the image size of 144 in order not to extract the noise image features

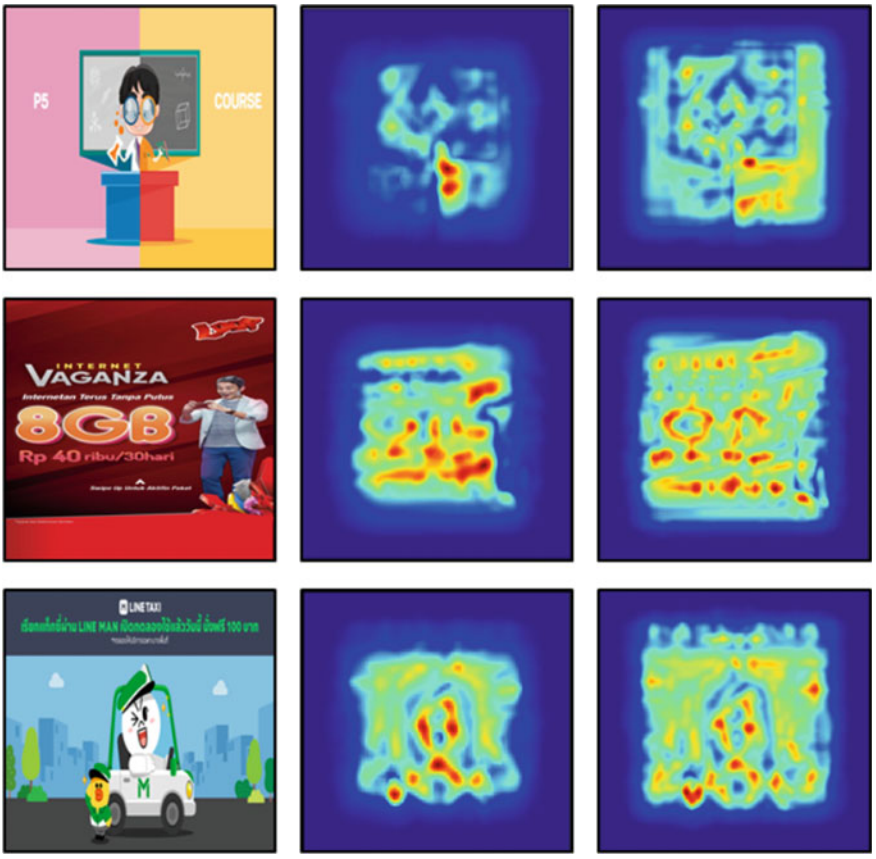


Fig. 4 The saliency map result after visualizing the activation matrix at the last convolutional layer of 2D CNN

immoderately in the applied Facebook dataset and able to achieve the lowest MSE of 0.68 by applying equation 12.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{12}$$

where the y_i indicates the actual CTR value while $\tilde{y}_i$ indicates the predicted CTR value through the proposed 2D CNN.

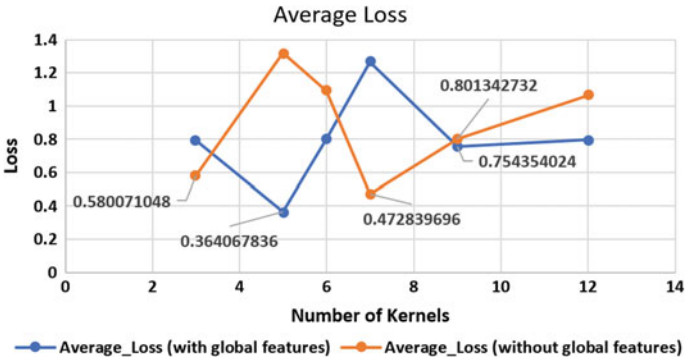


Fig. 5 The relationship of between the number of kernel and MSE Loss

4.3 1D CNN Experimental Results

In this section, the experiments of comparing the MSE between the click metrics data with global features, represented as C_{gf}, and without global features, C_{wgf} are conducted. First, the average loss result after applying the proposed 1D CNN on both metrics data as depicted in Fig. 5 will be discussed.

In Fig. 5, it is observed that the C_{gf} has achieved the lowest MSE of 0.364 when the number of kernels, K is 5. Subsequently, the average MSE begins to rise starting at $K = 6$. In contrast, C_{wgf} shows a continuous decrease of average MSE value starting from $K = 5$ and achieved the best average MSE value of 0.473 with $K = 7$. Nonetheless, C_{gf} and C_{wgf} show similar MSE value at $K = 9$. In overall, it is observed that with C_{gf}, it is able to achieve a low MSE value faster without requiring a larger kernel size like C_{wgf} in order to achieve a new lower MSE value from 0.580 at $K = 3$ to 0.473 at $K = 7$. Moreover, it also indicates that C_{gf} can achieve a much lower MSE as seen at $K = 5$. Therefore, to examine $K = 5$ is an optimum kernel size 1D CNN, the experiment of generating the standard deviation loss, Root Mean Square Error (RMSE), is computed to verify the proposed 1D CNN learning capability when $K = 5$ as shown in Fig. 6.

Figure 6 shows that the C_{wgf} at $K = 3$ has the lowest RMSE value of 0.153. However, recall that the MSE value of C_{wgf} at $K = 3$ is higher than the C_{gf} MSE value at $K = 5$. Although C_{wgf} at $K = 3$ shows a better learning capability compared to C_{gf} at $K = 5$ as it has a more rigid confidence interval, the natural characteristic of the metric data without the global features is not able to help in further reducing the MSE value. On the other hand, the second lower RMSE occurred at C_{gf} when $K = 5$ with the value of 0.189 has yield the lowest MSE. Therefore, it is proven that the proposed 1D CNN works best with $K = 5$ with the best MSE of 0.364 ± 0.189. Note that these average loss and standard deviation loss results for each number of kernels are generated and averaged by conducting the experiments for 5 times

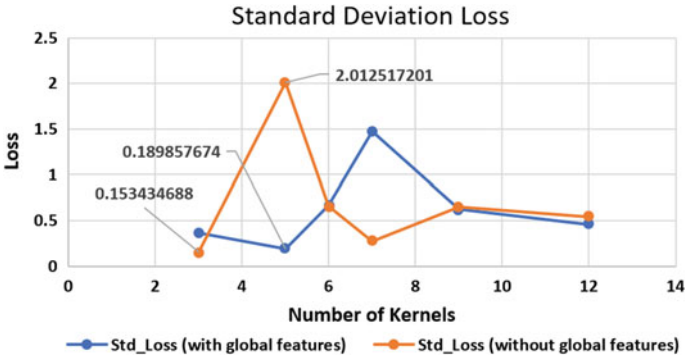


Fig. 6 The relationship between the number of kernels and RMSE

repeatedly as the stochastic learning in neural network will generate a different MSE value each time.

5 Conclusion

In this paper, a 2D CNN for extracting the salient features from the image and 1D CNN for predicting the CTR based on click metrics data were proposed. Furthermore, the experimental results showed that the loss of the proposed model can be further reduced after adding in the global image features. Nonetheless, it is noted that the experimental results showed in this paper were preliminary results as more display image data will be collected in the near future in order to verify and improve the performance of the proposed 2D CNN model. As for our future work, we will combine both CNN architecture models where the correlation between the feature maps of image and the corresponding click metric can be fused and thus, generating a more robust non-linear features to heighten the prediction of display CTR value.

References

1. Jones, J.P.: When Ads Work: new Proof that Advertising Triggers Sales. ME Sharpe (2006)
2. cisco.com: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper 2019. https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-738429.html#_Toc953325. Accessed 22 Mar 2019
3. Mei, T., Li, L., Hua, X.S., et al.: ImageSense: towards contextual image advertising. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **8**(1), 6 (2012)
4. Wang, J., Zhang, W., Yuan, S.: Display advertising with real-time bidding (RTB) and behavioural targeting. Found. Trends Inf. Retr. **11**(4–5), 297–435 (2017)

5. Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: estimating the click-through rate for new ads. In: Proceedings of the 16th International Conference on World Wide Web, pp. 521–530 (2007)
6. Liu, Q., Yu, F., Wu, S., et al.: A convolutional click prediction model. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1743–1746 (2015)
7. Guo, H., Tang, R., Ye, Y., et al.: DeepFM: a factorization-machine based neural network for CTR prediction. arXiv preprint arXiv:1703.04247 (2017)
8. Qu, Y., Cai, H., Ren, K., et al.: Product-based neural networks for user response prediction. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 1149–1154 (2016)
9. Chan, P.P., Hu, X., Zhao, L., et al.: Convolutional neural networks based click-through rate prediction with multiple feature sequences. In: IJCAI 2007–2013 (2018)
10. Zhou, G., Zhu, X., Song, C., et al.: Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1059–1068 (2018)
11. Cheng, H., Zwol, R.V., Azimi, J., et al.: Multimedia features for click prediction of new ads in display advertising. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 777–785 (2012)
12. Chen, J., Sun, B., Li, H., et al.: Deep CTR prediction in display advertising. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 811–820 (2016)
13. Dahlen, M.: Banner advertisements through a new lens. J. Advert. Res. **41**(4), 23–30 (2001)
14. Hasler, D., Suesstrunk, S.E.: Measuring colorfulness in natural images. In: Human vision and electronic imaging VIII, vol. 5007, pp. 87–96. International Society for Optics and Photonics (2003)
15. Huang, K.Q., Wang, Q., Wu, Z.Y.: Natural color image enhancement and evaluation algorithm based on human visual system. Comput. Vis. Image Underst. **103**(1), 52–63 (2006)
16. Azimi, J., Zhang, R., Zhou, Y., et al.: The impact of visual appearance on user response in online display advertising. In: Proceedings of the 21st International Conference on World Wide Web, pp. 457–458 (2012)

Functional Reactive EDSL with Asynchronous Execution for Resource-Constrained Embedded Systems

Sheng Wang and Takuo Watanabe

Abstract This paper presents a functional reactive embedded domain-specific language (EDSL) for resource-constrained embedded systems and its efficient execution method. In the language, time-varying values changes at discrete points of time rather than continuously. Combined with a mechanism to let users designate the update interval of values, it is possible to derive the minimal value-updates required to produce the user-desired output. Also, the event-driven backend asynchronously updates an input value when its value is required. In this way, we can greatly reduce the number of updates.

Keywords Functional reactive programming · Embedded domain-specific language · Embedded systems · Haskell

1 Introduction

A *reactive system* responds to external inputs in a timely manner. Robots that produce real-time motor commands according to continuously-changing environments and GUI systems whose contents change asynchronously with user inputs are two typical examples of reactive systems. In traditional sequential programming languages, we frequently use polling and/or event-driven callbacks to implement such systems. Unfortunately, these methods are complex and error-prone [3]. From the point of view of modularity, polling loops are not composable, and callbacks make control logic scattered across multiple pieces. What's worse, since the arrival of input events is unpredictable, programmers need to manage updates of mutable states carefully to preserve dependencies among variables.

S. Wang · T. Watanabe (✉)
Department of Computer Science, Tokyo Institute of Technology, W8-75, 2-12-1 Ookayama, Meguroku, Tokyo 152-8552, Japan
e-mail: takuo@acm.org

S. Wang
e-mail: kikyouer@gmail.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_12

Functional Reactive Programming (FRP) is a programming paradigm originated from dataflow languages. FRP introduces *time-varying values* (aka *signals*) to represent values that (continuously or discretely) change over time. Time-varying values can be composed and transformed as if they were plain values. By applying functional primitives such as `map`, `reduce` and `filter` on time-varying values, programmers can model the time-dependent relationships declaratively.

While FRP has been successfully applied in many areas such as computer animation and music composing, its usage on embedded systems is still limited. CFRP [14], Emfrp [13] and Juniper [9] tried to fill the gap by generating code with a small memory footprint. However, inefficiency still persists. These languages repetitively sample every input value and propagates update iteratively even if some input values do not contribute to the computation according to data dependencies. The unnecessary frequent activation of input sensors will significantly increase the battery consumption of the device.

The objective of our research is to propose an efficient FRP language for resource-constrained embedded systems. The runtime should be aware of unnecessary updates and automatically removes them from update iterations. We also aim to provide a language that is familiar to Haskell users and is easily extendable.

We present Hae, a code-generating embedded domain-specific language (EDSL) in Haskell. Instead of developing a standalone DSL, we use the technique called *deep embedding* [8]. A deeply embedded DSL overloads the host language's constructs and use them as combinators to construct the abstract syntax. Hae's user programs are written in Haskell source files and preprocessed by Haskell compiler. This process allows users to not only reuse all developing tools of Haskell but also utilize Haskell as a macro system for metaprogramming.

The FRP construct of Hae is different from that of languages such as CFRP or Emfrp. Time-varying values changes at discrete points of time rather than continuously. Combined with a mechanism to let users designate the update interval of values, we make it possible to derive the minimal value-updates required to produce the user-desired output. A Hae program is transformed to C++ code to be integrated with Hae's event-driven backend that asynchronously updates input values when they are required. In this way, we can greatly reduce the number of updates.

The rest of the paper is organized as follows. Section 2 briefly describes Hae using an example. Then, the execution model of the language with its optimization is discussed in Sect. 3. Section 4 describes the implementation of the language and Sect. 5 presents the evaluation result. Section 6 overviews related work and Sect. 7 concludes the paper.

```
import Hae.Expr
import Hae.Num
import Hae.Bool
import Hae.Compiler

tmp :: E Double -- temperature sensor input
tmp = input "tmp" (EveryS 1)

hum :: E Double -- humidity sensor input
hum = input "hum" (EveryS 1)

di :: E Double -- discomfort index
di = 0.81 * tmp + 0.01 * hum * (0.99 * tmp - 14.3) + 46.3

fan :: E Bool
fan = di .>= 75.0

main = putStrLn $ compile [OutputDef "fan" fan]
```

Listing 1 A simple fan controller in Hae

2 Language Hae

2.1 Overview

Hae[1] is an embedded domain-specific language (EDSL) in Haskell that generate C++ code to be integrated with different embedded developing tools. Hae provides users with essential FRP primitives and combinators to compose a static *signal graph*. A signal graph is a directed acyclic graph whose nodes and edges are time-varying values and their dependencies. Hae compiler translates the signal graph to C++ representation, which users can link against different event-driven backends such as Mbed OS.[2] Hae also provides meta-level libraries such as a static Vector library upon basic language elements. The data structure provided by these libraries exist only at compile-time.

A simple implementation of a fan controller in Hae is shown in Listing 1. We declare two input signals `tmp` and `hum` of type `Double` to represent the readings of temperature and humidity sensors respectively. The discomfort index[3] is computed from the current temperature and humidity. If the value is larger than 75.0, the value of signal `fan` will be `True`. This means that the fan is turned on while the air environment is uncomfortable. Users need to fill in I/O code in C++ for input (`tmp` and `hum`) and output (`fan`) nodes. After compilation, signals will be transformed to

[1] https://github.com/psg-titech/hae.

[2] https://www.mbed.com/en/platform/mbed-os/.

[3] A kind of human stress indicators. It is empirically known that 50% of people feel uncomfortable if it reaches 75.

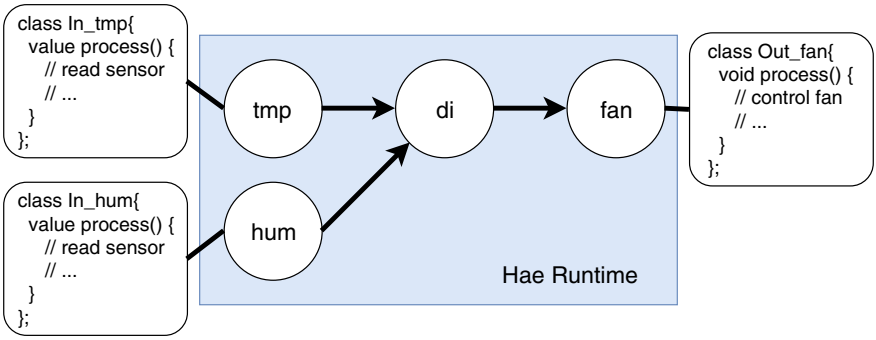


Fig. 1 Fan controller after compilation

nodes in Hae's C++ runtime, as depicted in Fig. 1. As we can see, Hae is a purely functional language where side effects can only happen within I/O nodes.

2.2 *Deep Embedding*

We claim that the code-generating EDSL in Haskell fits with FRP on resource-constrained systems very well in the following viewpoints:

- It enables us to reuse Haskell's modules system, type-checker, debugger and other tools, enjoying not only the ease of implementation but also a familiar development environment.
- We can implement functional language features such as higher-order functions and currying effortlessly. The host language will expand usages of these features at Hae's compile-time so that there is no runtime penalty.
- By embedding our language in Haskell, we automatically obtain a powerful and hygienic macro system. It is extremely useful for resource-constrained systems. For example, the `Vector` library built upon the macro system is able to fuse producer and consumer functions and eliminate all intermediate results. Such libraries are also easy to use as we can distinguish them from plain functions by their type signatures. This is further discussed in Sect. 2.5.

2.3 *Discrete Signals*

FRP with everchanging continuous signals is easy to understand and elegant: programmers assume that the system runs infinitely fast, and the precision depends merely on iteration interval chosen by the runtime. This may be the best for computer animation that FRP is originally designed for. But when it comes to resource-

```
a :: E Double a = 0.5

x :: E Int
x = input "x" (EveryMs 10)

y :: E Bool
y = (3 * x + 1) .> 5
```

Listing 2 Example definitions of signals

constrained hardware, we have to take efficiency into serious consideration. Iteration that is too fast causes high power consumption, and slow iteration makes the system unresponsive. As the continuous semantic forces the whole system to iterate at the same speed, it is difficult to choose a proper rate.

Event-driven discrete signals allow finer control over the updating process. Different signals can update asynchronously. We leveraged this feature of discrete FRP and designed a mechanism to perform updates only when they are needed. The mechanism is one of the main contributions of our work and is discussed in detail in Sect. 3.3.

Another advantage of discrete signals is that it is easier to incorporate asynchronous computation. If some computation is slow and we do not need the latest result of it, we can disconnect it from the main signal graph and use asynchronous signals to connect the input and output of that computation back to the system.

2.4 Signal Definitions

Embedded in Haskell, Hae shares the basic syntax and type system with its host language. As a result, Hae inherits the type definition, function definition and expressions like let-binding directly from Haskell. In this and next subsections, we will introduce how signals and other reactive primitives fit with Haskell.

We use type constructor `E` to denote a signal. The code snippet in Listing 2 shows definitions of a constant signal, an input signal, and a Boolean signal whose value is obtained by basic arithmetic operations.

As we can see, all literals are automatically lifted to constant signals. Currently, Hae has four primitive types: `Int`, `Float`, `Double`, and `Bool`. They can be compiled to their C++ counterparts. Higher-order signal (signal of signals) is not allowed for the sake of static construction of signal graphs.

The definition of input signal `x` is straightforward using function `input`. The first string parameter is used for Hae compiler to generate a function stub of the same name.

```
compile :: [OutputDef] -> String

-- to print out the translated string
main = putStrLn $ compile [OutputDef "out1" out1, ...]
```

Listing 3 Designating output signals

```
computeDi :: E Double -> E Double -> E Double
computeDi t h = t - 0.55 * (1 - 0.01 * h) * (t - 14.5)

di = computeDi tmp hum

curriedDi :: E Double -> E Double
curriedDi = computeDi tmp

computeDi' :: E (Double -> Double -> Double)
computeDi' = lift2 computeDi

-- operator |$| is required to apply a lifted function
di' = computeDi' |$| (tmp, hum)
```

Listing 4 Two types of functions

Finally, to complete a Hae program, we need to tell the compiler which output nodes we want to compile. As shown in Listing 3, function `compile` translates a list of output nodes to C++ code as a string.

2.5 Functions

Functions in Hae are categorized into two types. One has type $\mathtt{E}\,\alpha_1 \to \mathtt{E}\,\alpha_2 \to \cdots \to \mathtt{E}\,\alpha_m$ and the other has type $\mathtt{E}\,(\alpha_1 \to \alpha_2 \to \cdots \to \alpha_n)$ where types α_i do not contain type constructor `E`. The first type is essentially a Haskell function that any application of it will be inline-expanded. It also supports currying and higher-order functions. Besides inlined functions, it can also be used as a metaprogramming tool.

The second type of function actually exists at runtime as a node in the signal graph. It is better to use this kind of function if it will be called many times to reduce memory usage. The code in Listing 4 shows examples of usage and conversion from the first to the second type of functions.

The second type of function can be obtained by *lifting* a function of the first type. Users of existing FRP languages may have noticed that the type signature of our `lift` function is different from that of other languages. For example, `lift2` in the code snippet has type

$$\mathtt{lift2}::(E\ a \to E\ b \to E\ c) \to E(a \to b \to c)$$

```
data Vector a = Indexed { length :: Int, index :: Int -> a }
type EVector a = Vector (E a)

map :: (a -> b) -> Vector a -> Vector b
map f (Indexed l ixf) = Indexed l (f . ixf)

take :: Int -> Vector a -> Vector a
take n (Indexed l ixf) = Indexed (min n l) ixf

drop :: Int -> Vector a -> Vector a
drop n (Indexed l ixf) =
    Indexed (max 0 (l - n)) (\x -> ixf (x + n))

(...) :: Int -> Int -> EVector Int
(...) m n = Indexed (n - m + 1) (+ m)
```

Listing 5 Definition of `Vector` library

while its counterpart in CFRP would be typed as

$$\texttt{lift2}::(a \rightarrow b \rightarrow c) \rightarrow (E\ a \rightarrow E\ b \rightarrow E\ c).$$

The difference comes from the fact that Hae is an embedded language. Primitive data types (a, b and c) and functions (in the form of $a \rightarrow b$) of the host language cannot be directly reified without wrapping. Having to write `E` everywhere may seem tiresome, but it also gives a clear distinction between elements in Haskell and our DSL.

2.6 Datatype and the Vector Library

One limitation of deeply embedded DSL is that users cannot customize datatypes of the DSL without modifying the interpreter (compiler). A well-known technique to deal with this issue is to construct a small core language consisting of essential datatypes and build meta-level libraries of datatypes upon it using the host language Haskell. The design of Hae language takes the same approach. Here we briefly introduce how such a library can be built to show the expressiveness of Hae language.

Hae comes with a simplified version of Vector library in Feldspar [2]. A vector is like an indexed array except that it does not exist in memory at runtime. The definition of the vector type is shown in Listing 5. Data constructor `Vector` takes the length and the index function (function mapping an index to corresponding value stored in the vector) as parameters. The `map` function simply composes the provided function with the vector's index function. Other functions work similarly.

As we have discussed in Sect. 2.5, Haskell functions will be automatically inline-expanded. The producer function and consumer function of a vector will be fused together so that there will be no vectors of intermediate results.

Vector provides an efficient implementation of static-sized array in Hae. Other static container types can be easily built using the same method.

3 Execution Model

3.1 Updates on Signal Graph

The execution model of Hae is similar to that of Emfrp [13]. Signals and their dependencies form a directed acyclic graph. There are two ways to propagate updates on such a graph. They can be "pushed" from input nodes or "pulled" by output nodes. Pull-based FRP computes backward through the signal graph whenever results are demanded. As it is demand-driven, it eliminates unnecessary computation or polling on inputs. However, pull-based implementation requires lazy evaluation which limits its usage on resource-constrained systems.

Hae uses a push-based runtime. Comparing to pull-based FRP, it causes less latency between occurrence of an event and the reaction to it [6]. Updates are pushed following the order which is given by the topology sorting of the graph, from source nodes representing input signals to sink nodes representing output signals.

An important difference from Emfrp is that all input signals are not required to be updated in the same update iteration. In Hae, every signal contains information about its updating interval. Asynchronously, some signals update at a faster rate than others. In the iteration where fast signals update, slow signals do not need to activate. The runtime will push their previous value to its descendants. As a result, we no longer need to update the whole system whenever we get a new input.

We give users the ability to explicitly control the updating interval. When declaring input signals, users can specify update intervals of them. The update intervals of other signals that depend on the input signals can be calculated from them. For example,

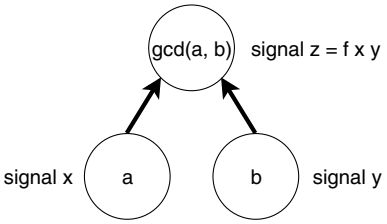


Fig. 2 Dependency

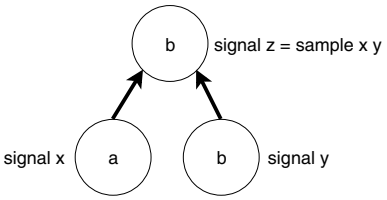


Fig. 3 Sampling

if signal `x` updates every 150 ms and signal `y` updates every 200 ms, then any signals that depend on these two signals will update every 50 ms (calculated by taking the GCD of the two), as shown in Fig. 2. We can also construct a new signal by sampling on another signals when its value changes. For example, in the case of `x` and `y`, the updating interval of `sample x y` is 200 ms, as shown in Fig. 3.

3.2 Stateful Computation

Hae provides two kinds of stateful computation. The first one is to refer to the previous value of a signal. This is implemented by inserting a `delay` node into the signal graph. The node will delay one event occurrence of a signal. If a signal emits `[0,1,2,3...]`, the delayed signal will yield `[0,0,1,2...]` in the corresponding update iteration.

The `@last` modifier in Emfrp serves a similar purpose of referring to previous values. But their semantics are different. Emfrp adopts the continuous signal model, `@last` in Emfrp refers to the value at the snapshot taken a moment ago (implemented as values in the last update cycle). Thus `a@last` and `b@last` are guaranteed to represent values taken in the same update cycle. In contrast, `delay a` and `delay b` in Hae do not necessarily represent values of the same time as updates are asynchronous.

Another kind of stateful computation is `foldp`. It behaves like the fold function in functional languages, but instead of accumulating list elements, it accumulates the history of values of a signal. `foldp` should be used with care since it cannot be optimized using the algorithm below.

3.3 Optimizing Push Timings

It is obvious that when we sample a fast-updating signal on a slower signal, all preceding signals of the sampled signal only need to update at the slower rate. Using this idea, we can calculate the actual update interval (which should be slower) by working from output signals backward the signal graph.

Consider the signal graph in Fig. 4. For signal `y` and `z` to update at interval b and c, signal `x` that provides value to them should update at both the intervals. Thus the actual update interval of signal `x` should be a list, containing those two update intervals. We apply the same method down the dependency graph and build the list of update intervals until reaching input signals.

Things become more complicated when stateful computations are introduced. If `foldp` is used, every previous value of preceding signals will contribute to the final accumulated value. In this case we cannot optimize the update interval.

For stateful computation consisting of `delay` nodes, we need to offset the update timing. Let us extend the denotation of update intervals to offsetted timing using

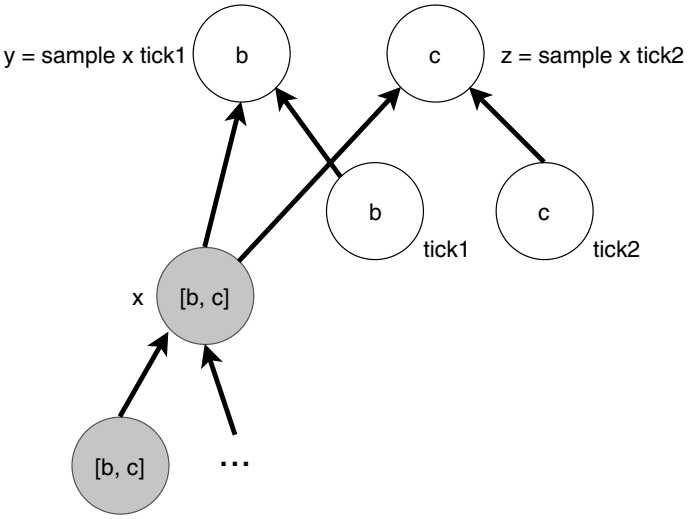


Fig. 4 Calculation of actual update interval

`interval(offset)`. For example, offsetted timing `60 ms(−1 ms)` means that the signal's value will be required at 59 ms, 119 ms,... With this extension, the actual required timing of a signal `actual(x)` can be computed by the following rules:

- Let `dep(x)` be signals that depend on `x`, `int(x)` be the user-designated update interval of `x`.
- If `x` is an output signal, `actual(x) = [int(x)]`.
- If `x` is a delayed signal, for each signal `s` in `dep(x)`, offset each timing in `actual(s)` by `-int(x)` and concatenate them together. For example, if `int(x) = 1 ms`, `dep(x) = [y, z]`, `actual(y) = [15 ms, 20 ms]` and `actual(z) = [50 ms(−1 ms)]`, then after offsetting and concatenating, `actual(x) = [15 ms(−1 ms), 20 ms(−1 ms), 50 ms(−2 ms)]`.
- If `x` is not delayed, we can just concatenate the timings without offsetting them.

4 Implementation

Hae is a code-generating embedded domain specific language. We will show the details of our prototype implementation in this chapter. The construction of a domain specific language within Haskell is explained in Sect. 4.1. Implementation of the C++ runtime is shown in Sect. 4.2. Finally, we introduce Hae's compiler in Sect. 4.3 with emphasis on a new optimization technique.

```
type E a = forall f. OE f a
newtype OE f a = OE (Ref (OpenExpr f a))

data OpenExpr (f :: * -> *) t where
  Inp :: HType t => String -> TimingDef -> OpenExpr f t
  Lit :: HType t => t -> OpenExpr f t
  PrimOp :: (Typeable a, Typeable b) =>
            PrimOpId -> (a -> b) -> OpenExpr f (a -> b)
  IfThenElse :: (Typeable t) => OE f Bool -> OE f t -> OE f t -> OpenExpr f t

  Var :: f t -> OpenExpr f t
  Lam :: (Typeable a, Typeable b) => (f a -> OE f b) -> OpenExpr f (a -> b)
  App :: (Typeable a) => OE f (a -> b) -> OE f a -> OpenExpr f b
  LetRec
    :: (CList ts, Typeable t)
    => (TList f ts -> TList (OE f) ts)
    -> (TList f ts -> OE f t)
    -> OpenExpr f t

  SampleOnChange :: Typeable a => OE f a -> OE f b -> OpenExpr f a
  Delay :: Typeable t => OE f t -> OpenExpr f t
  Foldp :: (Typeable a, Typeable b) => OE f (a -> b) -> OpenExpr f (a -> b)
```

Listing 6 Definition of core expression `OpenExpr`

4.1 EDSL Frontend

The core expression of Hae revolves around type `OpenExpr`. Type `E a`, which represents a signal of type `a`, is just a synonym of type `forall f. (Ref (OpenExpr f a))`. Before explaining the wrapper type `Ref`, let us first show the definition of `OpenExpr` in Listing 6. Some typical constructors of `OpenExpr` are also introduced below.

`OpenExpr` is the basic building block of *parametric higher-order abstract syntax* [4]. The first type parameter `f` is used to capture the type of shared expressions in `LetRec` bindings using the technique introduced by Oliveira et al. [12]. This technique ensures the type safety of the `LetRec` binder by statically encoding the types of binded expressions in *typed lists*. The Haskell compiler will then be able to reject illegal `LetRec` expressions.

Type constraint `HType t =>` on the `Inp` and `Lit` constructors ensures that we can only define input signals and constant signals of primitive types in Hae (`Int`, `Double`, `Float` and `Bool`).

Predefined primitive functions are reified by `PrimOp` constructor. By making `OE` an instance of `Num` type class, we can let Haskell implicitly insert corresponding `PrimOp` constructors when users write literal values and arithmetic operations. As a result, programming with signals feels no different than programming with plain values.

For user-defined functions that exist at run time (the second kind of function introduced in Sect. 2.5), we use `Lam`, `Var` and `App` constructors to reify them. `Lam` records the original unlisted function to be given a unique name when compiled.

```
data Ref a = Ref { refId :: Unique, deref :: a }

instance Eq (Ref a) where
  Ref x _ == Ref y _ = x == y
```

Listing 7 Definition of `Ref`

`Var` serves as a dummy argument to capture argument binding of that function. The application of these functions are represented by using `App`.

Finally, we have individual constructors for FRP primitives `SampleOnChange`, `Delay` and `Foldp` to distinct them from other expressions.

Careful readers may notice that we use type `OE` rather then `OpenExpr` in the recursive constructors. `OE` wraps `OpenExpr` with `Ref`, a data type that enables observable sharing [5] in a purely functional language.

Observable sharing is essential for an embedded domain specific language. Without observable sharing, our compiler cannot discover links among multiple references of a same value. As a result, for example, in expression `C*D*(D*D+D*E)`, `D` will be computed four times.

The definition of `Ref` is shown in Listing 7. It tags our value with a `Unique` identifier. `Ref` adds one level of indirection to raw values so that we can rediscover the sharing when analyzing these wrapped values by comparing their `refId`.

In Hae, users use smart constructors which wraps `Ref` automatically. For example, the smart constructor of `Delay` is defined as:

```
delay :: OE f a -> OE f a
delay = OE . ref . Delay
```

where `ref` is the function responsible to generate a unique `refId`. For simplicity, Hae implements `ref` function using `unsafePerformIO` in Haskell, as shown in Listing 8.

```
import Data.Unique (newUnique)

ref :: a -> Ref a
ref x = unsafePerformIO $ do
          u <- newUnique
          return (Ref u x)
```

Listing 8 Implementation of `ref`

4.2 Runtime

Hae uses a modified version of CFRP's event-driven runtime [14]. They share the same queue-based iteration logic within an update iteration:

1. During each update iteration, a node maintains the number of events it expects to receive.
2. When beginning an update iteration, input nodes are pushed to the update queue.
3. For each node `n` in the update queue, repeat the process until the queue is empty:
   a. `n.process()` is called to generate an update event. The actual computation of the node is done here
   b. The update event is sent to all of its children
   c. If a children has received enough events, it will be added to the update queue
   d. Finally, `n` is removed from the update queue

In CFRP, this update iteration is repeatedly executed in an infinite loop. To adapt it to our asynchronous execution model, we enhance input nodes with structure `Timing` (Listing 9). Every input node now contains the information of multiple timings. The runtime engine only initiates an update iteration at these timings. In the update iteration, input nodes which are not the initiators of the update will remain deactivated, sending empty update events to its children. After an update iteration is finished, the runtime engine schedules the next activation of these input nodes.

```
struct Timing {
    ttime period; // 64 bit int, the actual update interval
    std::vector<ttime> offsets;
    Timing(...) {...} // constructor
};
```

Listing 9 The data structure `Timing`

Note that `delay` is also a new type of node in Hae. Its implementation is straightforward. When a `delay` node is activated, it saves its newly received value to member variable `prev_` to be used as next iteration's output.

4.3 Compiler and Optimization

The compiling process of an EDSL in Haskell is slightly different from that of a standalone functional language. There is no need for lexing, parsing, α-conversion and even K-normalization as the host language automatically expands let bindings before constructing the DSL's abstract syntax.

We can also let Haskell do type inference for us by using `Data.Typeable` in Haskell's basic libraries. The `Typeable` class associates type representations to types. By deriving `OpenExpr` as an instance of `Typeable`. We can retrieve the type representation of any DSL expression using

$$\texttt{typeOf::forall}\ a.\texttt{Typeable}\ a \Rightarrow\ a \rightarrow \texttt{TypeRep}.$$

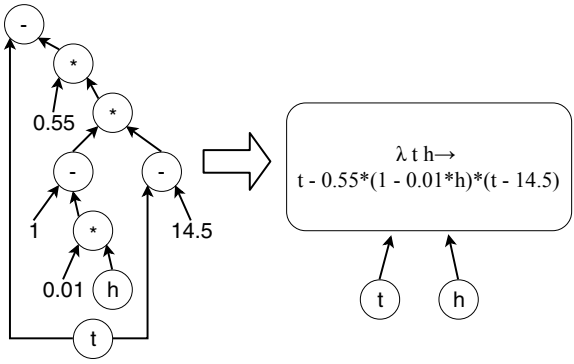


Fig. 5 A simple example of node merging

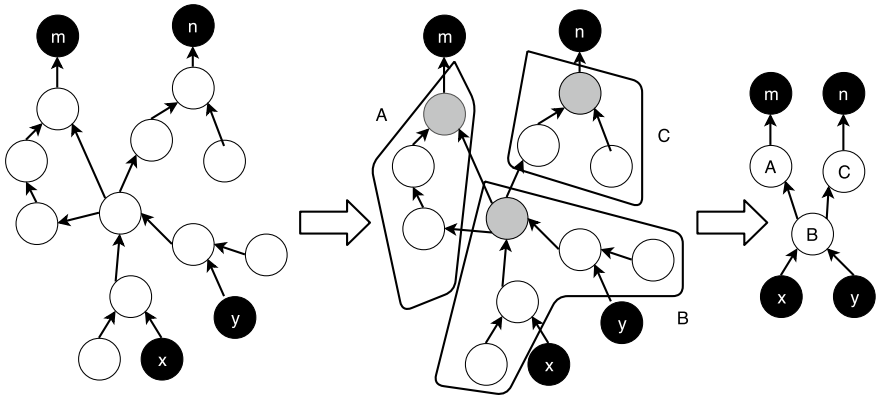


Fig. 6 A more general case of node merging

Since Hae generates C++ code, we can let C++ compilers handle most of the optimizations. However, there is still one essential optimization to do: merging adjacent primitive operations into a single node. Without merging, every arithmetic operation becomes an individual node at runtime, which is unacceptable considering time and space overhead.

Figure 5 demonstrates node merging on the computation of comfort index. Nodes of primitive functions and constants are merged into a single node that takes its input from two nodes.

A more general case is shown in Fig. 6. A white node denote a primitive or constant node. Other nodes that cannot be merged are marked black. Our node-merging algorithm partitions white nodes into 3 clusters. Each cluster contains exactly one *output node* which is marked gray. The output node must be the sink node within the partition. Finally, nodes in each partition are merged together, forming a simplified signal graph.

The key problem here is how to find the partition. Globally we maintain a map from node to cluster to record the membership of every white node. At the beginning, the map is empty. Then, for each output node, we traverse backward the dependency graph until a white node x that belongs to no cluster is found. This is the gray node of a new cluster. Then we start the subprocess below:

1. Mark node x gray.
2. Use node x as the source node for a new round of DFS. Note that we search only white nodes in this round of DFS.
3. During the DFS, set white nodes that previously belongs to no cluster as a member of cluster x. However, if we encountered a white node n that is already a member of another partition, it means cluster x depends on the value of n while n is not an output node. This is not allowed. In this case, we recursively start a new subprocess with $x = n$, effectively making a new partition from n.

The algorithm stops when every white node belongs to a partition.

In Emfrp, the partition of primitive nodes is explicitly defined by the user using keyword `node`. Hae makes this process automatic and optimal. Users can write expressions in whatever way they want without worrying about the overhead of additional intermediate nodes.

5 Evaluation

In this section, we evaluate the primary objective of our research—execution efficiency by two metrics. One is the number of times input sensors are activated. The other is the total number of node updates. Both measured in a fixed period of time. We run a test program using both Hae's asynchronous runtime and traditional repetitive iteration and compare the results of the two metrics. For evaluation purpose, we have built a simple repetitive execution engine upon Hae's asynchronous runtime (Listing 10).

```
void callback_(ttime now) {
  now_ = now;
  iterate_(); // the actual update iteration
  // call me after ITERATE_INTERVAL
  impl_set_cb_(now, ITERATE_INTERVAL);
  impl_yield_();
}
```

Listing 10 Repetitive execution

The test program (Listing 11) is a fan controller that switches on and off every minute depending on the sensor readings of temperature and humidity sensor. We

```
tmp :: E Double
tmp = input "tmp" (EveryMs 100)
hum :: E Double hum = input "hum" (EveryMs 1000)

avg3 :: E a -> E a
avg3 s = (s + prevS + delay prevS) / 3
  where
    prevS = delay s

tmpAvg = avg3 tmp

computeDi :: E Double -> E Double -> E Double
computeDi t h =
  0.81 * t + 0.01 * h * (0.99 * t - 14.3) + 46.3

di =
  sampleOnChange (compute_di tmpAvg hum) (fps (EveryS 5))

diAvg = avg3 di

fan :: E Bool
fan =
  sampleOnChange diAvg (fps (EveryMin 1)) .>= threshold
  where
    threshold = 75.0

main = putStrLn $ compile [OutputDef "fan" fan]
```

Listing 11 The test program in Hae

set the sampling interval of the temperature sensor `tmp` to 100 ms and the humidity sensor `hum` to 1000 ms. These values provide a baseline for computation of actual update timings. `avg3` is a helper function that utilizes Hae's discrete-time semantic to compute the average of the latest three values of a signal. We use it to smooth readings of `tmp` sensor. The discomfort index `di` is sampled every 5 seconds using built-in signal generator `fps`. Then, we take the average of discomfort index and use it to determine the switch of the fan once per minute.

Listing 12 shows the two input nodes after compilation using GHC 8.0.2. We can see that `tmp` contains 9 timings and `hum` have 3. During an update iteration, the runtime has to remember which node has initiated this iteration and also when to schedule the next round. Thankfully, the memory usage is linear to the total number of update timings.

We used a Linux-based backend to simulate running the test program for 60 min. The code generated by Hae and the backend are compiled by GCC 8.1. We ran the simulation on our PC running Fedora 28 on Intel's Core i7-6700K 4.0GHz with 16 gigabytes of RAM. The backend records the number of times of input node activations and node updates (by counting how many times a node's `process()` method is called). The repetitive reference runtime iterates at the rate of 100 ms (the same as

```
// ...
In_tmp n33;
n33.add_timing(hae::Timing(60000, 0, 0));
hae::ttime t33_1[8] = {-10200, -10100, -10000,
                       -5200, -5100, -5000, -200, -100};
n33.add_timing(hae::Timing(60000, 8, t33_1));
engine.register_input_node(&n33);
// ...
In_hum n47;
n47.add_timing(hae::Timing(60000, 0, 0));
hae::ttime t47_1[2] = {-10000, -5000};
n47.add_timing(hae::Timing(60000, 2, t47_1));
engine.register_input_node(&n47);
// ...
```

Listing 12 Input nodes after compilation

Table 1 Comparison of efficiency (in number of times)

	Iterations	Input activations	Node updates
Repetitive	36000	72000	648000
Async Exec	540	722	4509

the update interval of input signal `tmp`). Results are shown in Table 1. Both metrics are reduced by orders of magnitudes.

For Hae's asynchronous runtime, increasing the final sampling of `fan` from 60s to 120s will result in both the number of times of input activations and node updates in Hae's runtime being in half. Increasing the update interval of `tmp` or `hum` causes no change to the result because the number of timings in the test program is determined by the sample node. In contrast, for the repetitive runtime, making the final sampling slower makes no difference. But increasing the update interval of input signals effectively means we can iterate at a slower rate, resulting in fewer updates. To summary up, compared to traditional languages, Hae's runtime perform better when output signals operate slower and when input sensors run at a faster rate.

6 Related Work

6.1 FRP Languages for Small-Scale Systems

Flask [11] is a continuous FRP language targeting sensor networks. The two-stage language design separates the meta-language describing sensor network structure and the node-level language that compiles to NesC, a C-like language for sensor nodes. The runtime deploys node-level code to each sensors and constructs the network.

Native NesC code can be embedded in Flask by using Haskell's quasiquotation language extension.

Emfrp [13] achieved static memory footprint as a purely functional reactive language. An Emfrp program can be directly mapped to a static directed graph, eliminating recursion and any dynamic allocation of memory. Signals (called Nodes) in Emfrp are not first-class citizens. To reference a node one must supply with the name of it. The graph-like program is then transformed to C/C++ codes of the target platform with stub I/O functions for input and output nodes. Programmers fill in these stubs to connect the reactive part to other parts of the system.

Juniper [9] is a ML-like FRP language targeting the Arduino platform. Unlike Emfrp which limits the language's expressive power, Juniper is equipped with advanced language features such as anonymous functions and parametric polymorphism. The signal graph is dynamic in Juniper, allowing use of higher-order signals. Juniper is not a purely functional language.

6.2 Code-Generating Embedded DSL

There are two flavors of domain-specific language. A DSL can be either first-class, with its own compiler or interpreter, or embedded in a host language. The embedded approach has advantages in that it can utilize the host language's syntax and ecosystem.

Haskell is a functional language with powerful type system and overloading capabilities. Previous research has explored Haskell's ability to not only run DSLs in its own runtime, but also generate code that can be interpreted by external programs. Code-generating embedded DSL leverages both the syntactic convenience of the host language and the flexibility of a customized runtime.

Elliott *et al.* demonstrated techniques to capture and reify variable bindings in Haskell in their image-processing language Pan [7]. Value sharing and recursion in EDSLs is made possible by using I/O based *observable sharing* [5] or *parametric higher-order abstract syntax* [4]. Some languages, such as [1, 2, 10], are built upon these techniques to generate all kinds of codes from CUDA, DSP algorithms to even Haskell code itself.

7 Conclusion

We have designed and implemented Hae, a functional reactive programming language targeting small-scale embedded systems. Hae showed that we can rule out unnecessary updates by combining the discrete-time semantic and explicit update intervals. This makes Hae much more efficient than previous FRP languages for embedded devices.

The design choice of a code-generating embedded domain specific language let us take advantage of a familiar developing environment, a powerful macro platform, flexible choice of backends and ease of implementation. The algorithm we used for merging primitive nodes is not only essential for embedded DSLs but also useful for standalone FRP languages.

Hae is currently a prototype language that requires polishing. Some designs may not be the best choice for embedded development. More case studies about FRP and embedded programming need to be done, especially those that may benefit from meta-programming. The survey will not only inspire new ideas but also provide better ways to evaluate our work.

As for language features, the ability to change update intervals at runtime can be an interesting research direction. While this will require more information about signal nodes to be maintained at runtime, the flexibility it brings should outweigh the cost.

Acknowledgements This work is supported in part by JSPS KAKENHI Grant No. 18K11236.

References

1. Ankner, J., Svenningsson, J.: An EDSL approach to high performance Haskell programming. In: ACM SIGPLAN Symposium on Haskell (Haksell 2013), pp. 1–12. ACM (2013). https://doi.org/10.1145/2503778.2503789
2. Axelsson, E., Claessen, K., Sheeran, M., Svenningsson, J., Engdal, D., Persson, A.: The design and implementation of Feldspar: an embedded language for digital signal processing. In: IFL 2010: Implementation and Application of Functional Languages. Lecture Notes in Computer Science, vol. 6647, pp. 121–136. Springer, Berlin (2010). https://doi.org/10.1007/978-3-642-24276-2_8
3. Bainomugisha, E., Carreton, A.L., Van Cutsem, T., Mostinckx, S., De Meuter, W.: A survey on reactive programming. ACM Comput. Surv. **45**(4), 52:1–52:34 (2013). https://doi.org/10.1145/2501654.2501666
4. Chlipala, A.: Parametric higher-order abstract syntax for mechanized semantics. In: 13th ACM SIGPLAN International Conference on Functional Programming (ICFP 2008), pp. 143–156. ACM (2008). https://doi.org/10.1145/1411204.1411226
5. Claessen, K., Sands, D.: Observable sharing for functional circuit description. In: Advances in Computing Science (ASIAN '99). Lecture Notes in Computer Science, vol. 1742, pp. 62–73. Springer, Berlin (1999). https://doi.org/10.1007/3-540-46674-6_7
6. Elliott, C.: Push-pull functional reactive programming. In: Proceedings of the 2nd ACM SIGPLAN Symposium on Haskell (Haskell 2009), pp. 25–36. ACM (2009). https://doi.org/10.1145/1596638.1596643
7. Elliott, C., Finne, S., de Moor, O.: Compiling embedded languages. J. Funct. Program. **13**(3), 455–481 (2003). https://doi.org/10.1017/S0956796802004574
8. Gill, A.: Domain-specific languages and code synthesis using Haskell. ACM Queue **12**(4) (2014). https://doi.org/10.1145/2611429.2617811
9. Helbling, C., Guyer, S.Z.: Juniper: A functional reactive programming language for the Arduino. In: 4th International Workshop on Functional Art, Music, Modelling, and Design (FARM 2016), pp. 8–16. ACM (2016). https://doi.org/10.1145/2975980.2975982

10. Mainland, G., Morrisett, G.: Nikola: Embedding compiled GPU functions in Haskell. In: 3rd ACM Symposium on Haskell (Haskell 2010), pp. 67–78. ACM (2010). https://doi.org/10.1145/1863523.1863533
11. Mainland, G., Morrisett, G., Welsh, M.: Flask: Staged functional programming for sensor networks. In: 13th ACM SIGPLAN International Conference on Functional Programming (ICFP 2008), pp. 335–346. ACM (2008). https://doi.org/10.1145/1411204.1411251
12. Oliveira, B.C.d.S., Löh, A.: Abstract syntax graphs for domain specific languages. In: Workshop on Partial Evaluation and Program Manipulation (PEPM 2013), pp. 87–96. ACM SIGPLAN, ACM (2013). https://doi.org/10.1145/2426890.2426909
13. Sawada, K., Watanabe, T.: Emfrp: A functional reactive programming language for small-scale embedded systems. In: MODULARITY Companion 2016: Companion Proceedings of the 15th International Conference on Modularity, pp. 36–44. ACM (2016). https://doi.org/10.1145/2892664.2892670
14. Suzuki, K., Nagayama, K., Sawada, K., Watanabe, T.: CFRP: A functional reactive programming language for small-scale embedded systems. In: Theory and Practice of Computation (Proceedings of WCTP 2016), pp. 1–13. World Scientific (2017). https://doi.org/10.1142/9789813234079_0001

Adaptive Midpoint Relay Selection: Enhancing Throughput in D2D Communications

Ushik Shrestha Khwakhali, Prapun Suksompong and Steven Gordon

Abstract D2D communications in cellular networks can leverage social information of users in relay selection to improve its performance. As a relay can be deployed to either extend the range or improve quality of D2D communications, social trust among users in the network can be exploited for selection of relay nodes to yield higher throughput. The analysis of results shows that midpoint relay selection scheme with social trust offers higher throughput compared to hybrid relay selection scheme. It is seen that the maximum of average throughput is significantly dependent on the lower limit of the support rather than how fast the distribution tail decays. Moreover, performance of the scheme is unsatisfactory when social trust among the nodes are high. We have proposed adaptive midpoint relay selection scheme that is designed to achieve the performance of MRSS-ST when the social trust among the nodes are low and achieve the performance of M-Nearest when the social trust are high.

1 Introduction

This section introduces device-to-device (D2D) communication technology, explains the problem of relay selection when a relay transmits at variable transmission power, and briefly describes different approaches taken for the selection of a relay to enhance the throughput of D2D communications.

U. Shrestha Khwakhali (✉)
Vincent Mary School of Engineering, Assumption University, Samut Prakan, Thailand
e-mail: ushik.shrestha@gmail.com

U. Shrestha Khwakhali · P. Suksompong
School of ICT, SIIT, Thammasat University, Bangkok, Thailand
e-mail: prapun@siit.tu.ac.th

S. Gordon
School of Engineering and Technology, CQUniversity, Rockhampton, Australia
e-mail: s.d.gordon@cqu.edu.au

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_13

1.1 D2D Communications

Device-to-device (D2D) communication technology enables direct communication between the devices in proximity. The spectrum used in mobile cellular networks can be used for D2D communications under the control of a base station (BS). However, unlike general mobile communication between devices in which all the user data goes through a BS, D2D communication uses cellular spectrum for the information exchange bypassing the BS. Furthermore, a relay can be used in D2D communications when the link is intermittent or need to extend the communication range [1–4], whenever possible. Therefore, D2D relay communication is considered an important feature in next generation mobile networks [5].

We have examined influence of social factor on the transmission power of a relay and analyzed its impact on the throughput of D2D communications. We assume that a relay forwards data of a node with a transmission power proportional to the social trust value that the relay has for the source. The nodes in a network are said to have relationship between them when the users of the nodes make calls or exchange information with each other. The social trust value between the pair of nodes can be calculated using the frequency and duration of calls made by the users [6, 7].

In this work, our focus is to select a relay that maximises the throughput of D2D communications. We have done extensive analysis of the Midpoint Relay Selection Scheme with Social Trust (MRSS-ST) proposed in [8]. Based upon shortcoming found from our analysis of MRSS-ST, we have proposed Adaptive Midpoint Relay Selection (Adaptive-MRS) scheme.

1.2 Problem in Relay Selection

In cooperative networks, a relay node can be used to forward data from a source to a destination to improve performance of the communication. Different approaches have been designed in the past for the selection of a relay that maximises throughput. The selection of a relay offering maximum throughput becomes more challenging when a relay can choose or adjust its transmission power level. The difficulty while selecting a relay is due to the fact that the throughput not only depends on the link lengths from the source to the relay and that from the relay to the destination, but also on the transmission power of the relay.

A relay (user) may be interested in transmitting at a higher power only for its friends, not for others. A higher transmission power of a relay increases average throughput of the D2D communication at the cost of increased power consumption at the relay. Therefore, a social relationship among users can be utilised for the selection of a relay node. Higher social trust may lead to higher SINR, potentially resulting to higher throughput. This is following the same assumption used in paper [6, 7].

The strength of relationship between people are usually bidirectional [9]. Therefore, we consider that the social trust value that a relay has for the source can be estimated by the social trust value the source has for the relay. In addition to the social trust value, channel conditions also affect the throughput. The channel conditions can be known by probing of nodes before the selection of a relay.

1.3 Related Works

The authors in [10] proposed a social-aware enhanced D2D communication architecture using social networking characteristics in system design. They did the qualitative analysis of the benefits from social features in D2D communications. They also quantified the achievable gain in social-aware D2D communications. Physical-social graphs created by the authors in [11] are used in cooperative D2D communications.

An optimal stopping approach for relay selection proposed in [7] is social trust based cooperative D2D relaying. It considers physical and social distances among users. Socially trusted nodes are sequentially probed in each timeslot before the relay selection. The drawback of this solution is that any node may be probed regardless of its location if the node is socially connected to the source.

Social information of users can be used to assist D2D communications in several ways. The relay transmission power is considered to be proportional to social trust between the relay and the source in [7]. Moreover, the power is assumed to be positive linear function of social relation between users in [6]. In [5], social relationship between people is used as an incentive factor to build distributed incentive mechanism. The assumption made by these researchers is that a mobile node relays data for other nodes at high transmission power when it has a strong social relationship with them. Social relationship developed between users is measured by the interaction between them. This is also the assumption in our work. Other researchers uses this concept of varying transmit power of a relay as social selfishness. They termed social selfishness as the tendency of nodes willingness to forward data for nodes in same social community and less interest to relay data for nodes outside social community [12].

Hybrid Relay Selection (HRS) scheme is proposed by Pan and Wang in [6]. In HRS, a source probes nodes within a circular region with the source at the center to select a relay having social trust above social threshold value. It is demonstrated that HRS provides significantly higher throughput compared to Distance-based Relay Selection and Social-based Relay Selection schemes. Moreover, the authors suggested relay selection region should neither be too small nor too large. The disadvantage of HRS is that, it does not have probe limitation, and nodes are probed around the source.

In Social-aware Midpoint Relay Selection Scheme (SMRSS) [13], and MRSS-ST [8], a relay is selected among the nodes that are located around midpoint of the distance between the source and the destination. The simulation results show that SMRSS and MRSS-ST both have higher performance than that of HRS in all the

considered scenarios. The probe limit in MRSS-ST further enhanced performance for large values of search radius compared to SMRSS.

1.4 Our Contributions

In this paper, we have done extensive performance analysis of MRSS-ST [8] and proposed Adaptive-MRS scheme to address the drawback of MRSS-ST. The main contribution of our work are as follows:

- We have performed the extensive analysis of MRSS-ST to show that MRSS-ST performs better than HRS in different social trust scenarios in networks having different node densities. However, we see that the performance of MRSS-ST is unsatisfactory compared to M-Nearest scheme when social trust among the nodes are high. In this work, the maximum of average throughput is used as a metric to compare performance of the schemes, which is different than in [8]. The significant findings of this work is to analyze performance variation of different relay selection schemes with the change in social trust scenarios.
- Based upon the insights from our extensive analysis, we have proposed Adaptive-MRS scheme. Adaptive-MRS is designed to select a relay nearest to midpoint when the social trust among the nodes are high, and probes nodes before the relay selection otherwise.

2 System Model

In this section we present the models of the network, communication links, and social trust that are assumed in our design and analysis.

2.1 Node and Network Model

Figure 1 shows the system model used in this work which is same as in [8]. It consists of a cooperative network where n_{total} mobile nodes held by people are under the coverage of a BS. The GPS in the nodes are assumed to be continuously updating their location to the BS. All idle nodes have potential to forward packets from the source s to the destination d as a relay r.

P_{max} is the maximum allowable transmission power used for a D2D communication which is set by the BS. The nodes in the network transmit at power $P_{i,j} \leq P_{max}$. A source node always transmits at the maximum allowable transmission power i.e. $P_{s,j} = P_{max}$. The transmission power of a relay $P_{r,d} \leq P_{max}$ is linearly proportional to the strength of social trust between it and the source, denoted by $\beta_{r,s}$ [6].

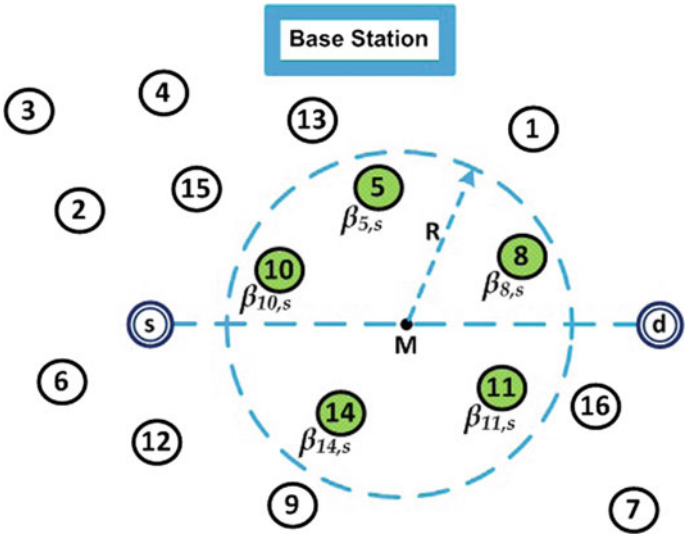


Fig. 1 MRSS-ST System Model

TDMA is used for the data transmission where a source can communicate with a destination either directly (denoted as a direct link) or via a relay (denoted as a D2D link). We consider the device relaying with operator controlled link establishment and direct D2D communication with operator controlled link establishment scenarios as envisioned in [9]. Decode and forward technique is used to relay data in D2D communications [6, 7, 12]. Now, we will present the communication link model used in this research for the throughput calculation.

2.2 Communication Link Model

The SNR of a direct link can be expressed as

$$\gamma_{s,d} = \frac{P_{s,d} D_{s,d}^{-\theta}}{N} \tag{1}$$

where $P_{s,d}$ is the transmission power of a signal from the source to the destination, $D_{s,d}$ is their distance, N is the noise power and θ is the pathloss exponent.

According to Shannon's Capacity formula, the data rate of a direct link is calculated as

$$C_{s,d} = B \log_2(1 + \gamma_{s,d}) \tag{2}$$

where B is the bandwidth of a channel.

The data rate of a D2D link having full duplex decode and forward relaying is given by

$$C_{s,r,d} = B \min\{\log_2(1 + \gamma_{s,r}), \log_2(1 + \gamma_{s,d} + \gamma_{r,d})\} \tag{3}$$

where $\gamma_{s,r}$ is SNR of the signal from the source to the relay, $\gamma_{s,d}$ is that from the source to the destination and $\gamma_{r,d}$ is that from the relay to the destination.

Taken into account the time wasted from probing, the throughput of a D2D communication is calculated as

$$T_{s,d} = \frac{C_{s,r,d}\{t - (\tau \times p)\}}{t} \tag{4}$$

where t is the timeslot duration, τ is the probe duration and p is the number of probes [6, 7, 11, 12].

2.3 Social Trust Model

Generally, social trust among majority of people in a community are weak. Only few people have high social trust between them. The authors in [6, 14] have modeled social trust using Pareto distribution to represent a heavy-tailed distribution. Thus social trust among people can be charaterised by a Pareto distribution which is defined as

$$f_X(x) = \frac{\alpha L^\alpha x^{-\alpha-1}}{1 - \left(\frac{L}{H}\right)^\alpha}, L \leq x \leq H, \text{and } \alpha > 0 \tag{5}$$

where α, L and H denote the shape, scale and upper limit parameters, respectively [15]. The value of α determines how fast the distribution tail decays and L is the lower limit of the support. Pareto distribution is used as a social trust model in this work. Different values of social trust among nodes are achieved by varying α and L, keeping H constant. Each node knows social trust value it has for other nodes.

Although MRSS-ST and MRSS with social distance (MRSS-SD) are two relay selection schemes proposed in [8], MRSS-ST has a superior performance compared to MRSS-SD when the BS is considered to adaptively adjust search radius. In next section, MRSS-ST is summarised and comparative analysis of performance of MRSS-ST along with other schemes is presented in the subsequent section.

3 Midpoint Relay Selection Scheme with Social Trust

In [8], a relay selection scheme is designed to select a socially trusted relay node located around the midpoint of the distance between the source and the destination. The social trust between users are utilised for the relay selection. The communica-

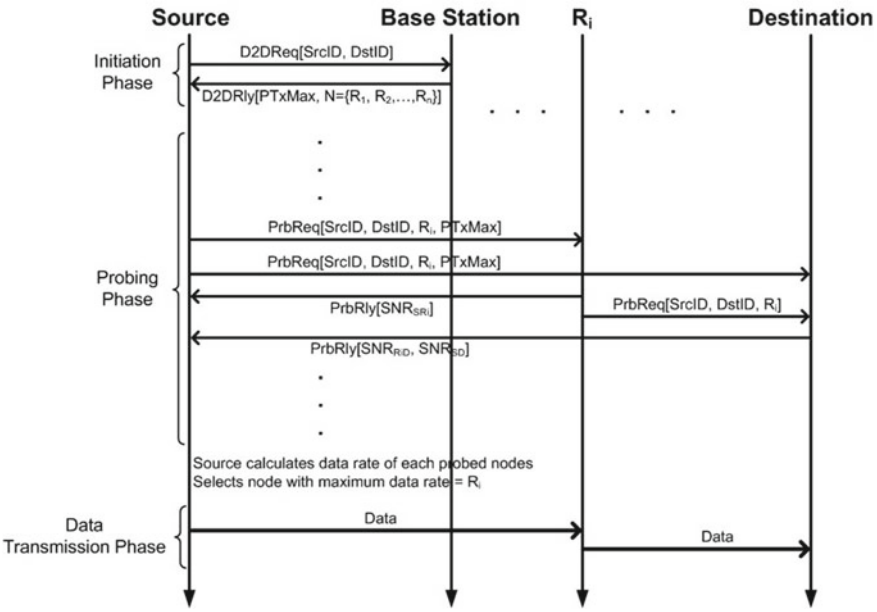


Fig. 2 Message flow diagram

tion protocol designed for the relay selection exchanges information between nodes during sequential probing as depicted in Fig. 2.

Algorithm 1 Commands performed at Base Station

// Given a source, destination and search radius, BS determines list of idle nodes within search radius and maximum allowable transmission power.
Input: s, d and R
Output: n_{circle} and P_{max}
1: Receive D2D request from s
2: Calculate Midpoint, m = Midpoint(s, d)
3: **for** $radius = 1 : R$ **do**
4: Identify n_{circle} = NodesWithinCircle(m, $radius$)
5: **end for**
6: Determine P_{max}
7: Send n_{circle} and P_{max} to s

Two messages are exchanged during the probe initiation phase. A source node requests for D2D communication by sending a message to the BS containing the source and the destination identity. In response to that, the BS determines idle nodes located around the midpoint of the distance between the source and the destination. The list of idle nodes are sent to the source by the BS together with maximum allowable transmission power as in Algorithm 1. Upon receiving the list, the source

filters nodes that have social trust value above social threshold which are known as potential relay nodes. Among those nodes, at most l nodes having maximum value of social trust are selected as candidate relay nodes as implemented in Algorithm 2.

Algorithm 2 Commands performed at Source

// A source determines whether to use a relay or not. If a relay is to be used, it filters nodes according to the input scheme. A relay is selected such that it maximises the throughput of D2D communication.
Input: s, d, n_{circle}, S_{thres}, P_{max}, l, τ and $scheme$
Output: $t_{s,r,d}$

if len(n_{circle}) = 0 **then**
  Calculate $C_{s,d}$ as in equation 2
  $t_{s,r,d} = C_{s,d}$
else
  Identify $n_{S_{thres}}$
  if len($n_{S_{thres}}$) = 0 **then**
    $t_{s,r,d} = C_{s,d}$
  else
    if len($n_{S_{thres}}$) > l **then**
      if $scheme$ == *MRSS-ST* **then**
        $n_{S_{thres}} = l$ nodes with maximum social trust values
      else
        $n_{S_{thres}} = l$ nodes nearest to midpoint
      end if
      Sequentially probe nodes in $n_{S_{thres}}$ and d
      Calculate $C_{s,r,d}$ through each of nodes in $n_{S_{thres}}$ as in equation 3
      Select maximum $C_{s,r,d} = max(C_{s,r,d})$
      if $max(C_{s,r,d}) > C_{s,d}$ **then**
        Relay is selected
        $t_{s,r,d} = max(C_{s,r,d}) \times (T - (\tau \times n_{S_{thres}}))$
      else
        $t_{s,r,d} = C_{s,d} \times (T - (\tau \times n_{S_{thres}}))$
      end if
    end if
  end if
end if

The source sequentially probes all the candidate relay nodes and destination. During the probing phase, four messages are exchanged per candidate relay node to learn about the link conditions. A candiate relay node sends SNR value of the received probing packet to the source and also forwards the packet to the destination as detailed in Algorithm 3. After receiving the probing packets from candidate relay node and source, the destination calculates SNR of each of those packets and sends the SNR values to the source as implemented in Algorithm 4. These values are used to calculate data rate offered by each of the nodes and data rate of direct transmission. MRSS-ST considers probing duration and either selects a node that offers maximum throughput as a relay or chooses direct data transmission. The user data transmission occurs during data transmission phase after determining whether to communicate via a relay or have direct communication. The detailed design of MRSS-ST is presented in [8].

Algorithm 3 Commands performed at Candidate Relay Nodes

// Given a source and a destination pair, candidate relay nodes forward probes to the destination. User data is also forwarded from the source to the destination and vice-versa.
Input: s, d, P_{max} and probe
Output: SNR values, relay data

```
Receive probe message from s
Send SNR value of probe received to s
Send probe to d
Send user data from s to d and vice-versa
```

Algorithm 4 Commands performed at Destination

// Receives probes and data sent to destination. It sends SNR values of the probes to the source as well as data.
Input: probe and data
Output: SNR values, data

```
Receive probe messages from s and r
Send SNR values of probe received from s and r to s
Receive data from r sent by s
Send user data to s via r
```

4 In-depth Analysis of MRSS-ST

In the section, we present the performance analysis of MRSS-ST along with other relay selection schemes. Adaptive-MRS is presented in the subsequent section.

4.1 Analysis Approach

We analyse the performance of different relay selection schemes. The comparative analysis of MRSS-ST is done against following schemes:

- HRS scheme, proposed in [6].
- M-Nearest scheme, where a relay is selected nearest to midpoint of the distance between the source and the destination.
- M-Nearest_MaxTx scheme, where a relay is selected nearest to midpoint of the distance between source and destination assuming relay always has maximum allowable transmission power. In reality, nearest to midpoint node does not always transmit at maximum power but the scheme is used as a benchmark scheme.

For the performance metric, we use the average throughput of each scheme. However to determine the average throughput, we do not compare them at the same search radius. The radius that is used is the radius that maximises the average throughput which we are going to refer as an optimal search radius. Such comparison makes more sense because each scheme can adaptively adjust the search radius to achieve its best performance. This is different from [8, 13] where the schemes are compared at the same search radius.

Table 1 Simulation setup

Simulation parameters	
Simulation software	Octave
Mobility model	Random waypoint
Network width	100 and 1000 m
Speed of nodes, S	0–2 m/s
Maximum pause time	5 s
Simulation period	1000 s
Channel bandwidth, B	1 MHz
Pathloss exponent, θ	4
Noise power, N	−114 dBm
Social trust, β	0–1
Social threshold, S_{thres}	0.3
Probe limit, l	10
Scale, L	0.001, 0.01, 0.1, 0.2, 0.5, 0.8, 0.99
Shape, α	1.001, 1.01, 1.1, 1.2, 1.3, 2, 5
Upper limit, H	10
Transmission power, $P_{s,d}$	20 dBm
Max. allowable tx power, $P_{s,j}$	20 dBm
Timeslot duration, t	1 s
Probe duration, τ	0.01 s
Search radius, R	Upto 500 m

4.2 Simulation Setup

Table 1 shows the details of the simulation setup used in our research. The nodes in the network are uniformly distributed over the network area. The social trust value between two nodes follows Pareto distribution. Different social trust scenarios are achieved by varing shape and scale parameters. The results based on the simulation setup is presented in the next section.

4.3 Results

In this section, we present the optimized average throughput that can be achieved from different relay selection schemes analysed under various social trust scenarios having different node densities.

Figure 3 shows the three dimensional comparison of maximum of average throughput offered by the different relay selection schemes in a square network having network width of 100 m. As expected, the throughput of M-Nearest_MaxTx has

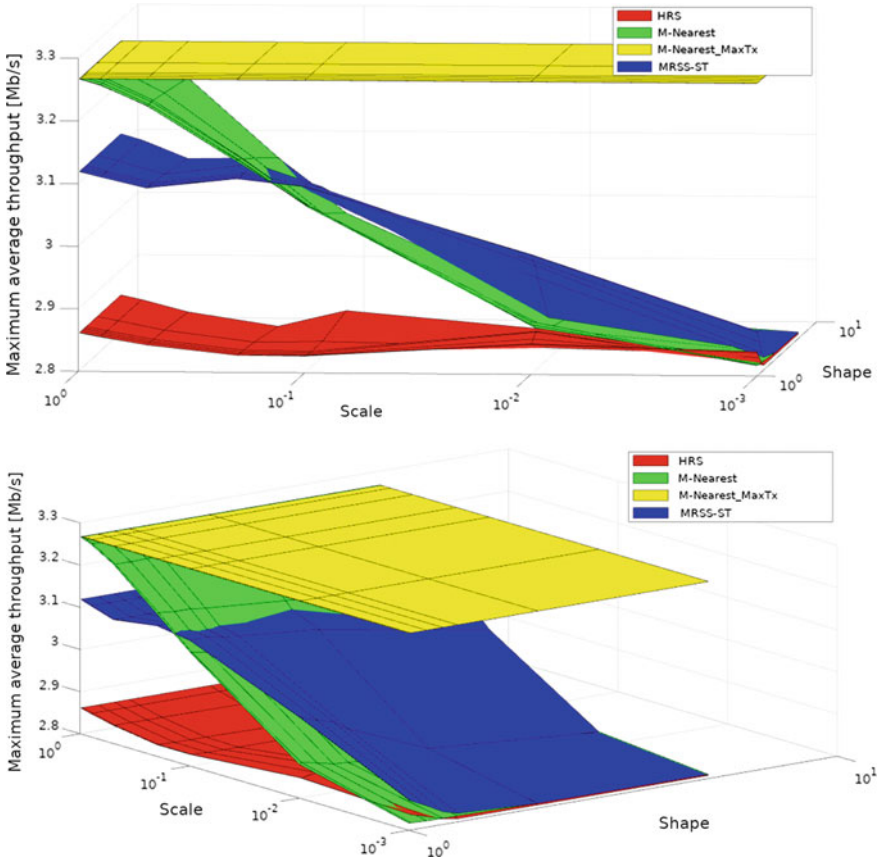


Fig. 3 Maximum of average throughput attained by different relay selection schemes for network width of 100 m (two different viewpoints of the same 3D plots)

the highest value among all the other schemes and remains constant for all values of shape and scale. For all the other schemes, the throughput increases with the increase in Scale value. Comparatively the throughput is less affected by the shape value. The throughput of MRSS-ST is significantly higher than that of HRS for all range of shape and scale. This is mainly because of the balance of link lengths between the source to the relay and from the relay to the destination in MRSS-ST. In addition, a probe limit also contributes to the performance enhancement of MRSS-ST in dense networks with high probability of social trust. This is because in such networks, number of nodes having social trust above social threshold can be large. Probing of all those nodes can consume large duration of a timeslot. As a result, the throughput decreases. However, with the use of probe limit, at most l nodes having maxium social trust values are only probed, thereby contributing to enhance the throughput.

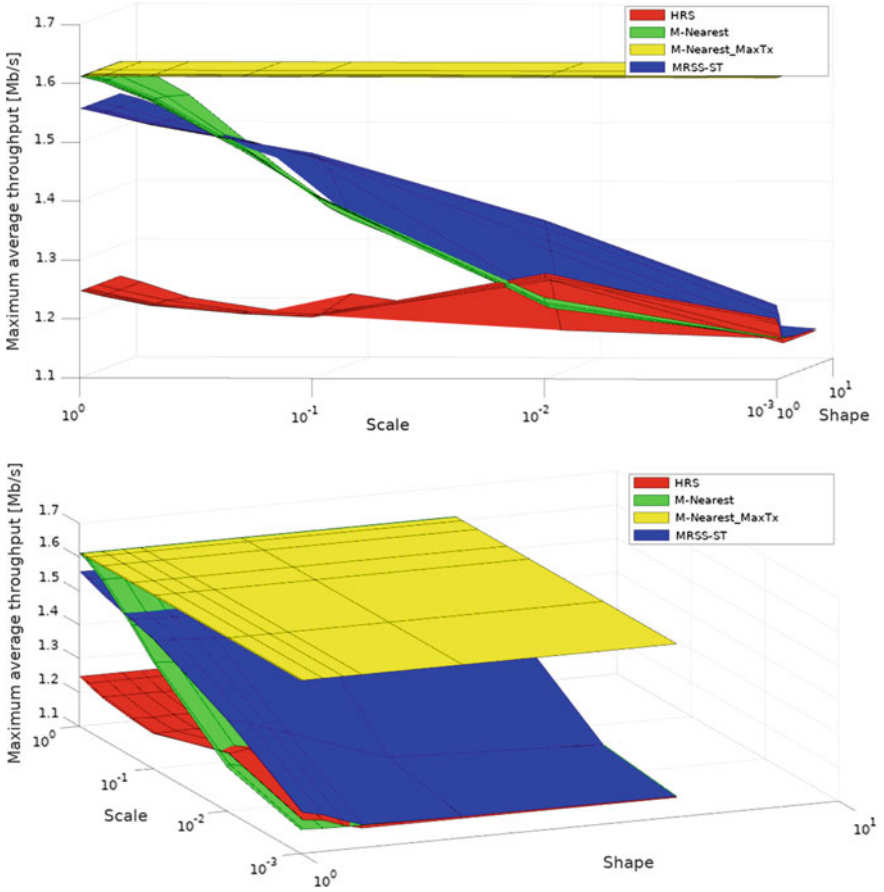


Fig. 4 Maximum of average throughput attained by different relay selection schemes for network width of 1000 m (two different viewpoints of the same 3D plots)

For the higher values of scale (approximately above 0.15), the throughput of M-Nearest is greater than that of MRSS-ST. When social trust among the nodes are very strong, all the nodes can potentially transmit at a high power. The probing of nodes before selection of a relay in MRSS-ST reduces time for actual data transmission. However, user data is directly transmitted in M-Nearest. This results in superior performance of M-Nearest compared to MRSS-ST for high value of scale.

Figure 4 shows the comparison of maximum of average throughput offered by the different relay selection schemes in a square network having network width of 1000 m. The trend of throughput variation for all the schemes are similar to that in network having width of 100 m. The comparison shows that Fig. 3 has higher value of the throughput than that of Fig. 4.

More importantly we also show that the maximum of average throughput of MRSS-ST increases significantly with the increase in scale value and is less dependent with variation of shape value in both the network scenarios. The insights obtained from the results motivated to develop Adaptive-MRS scheme presented in the next section.

5 Improvement Upon MRSS-ST: Adaptive-MRS

From careful analysis of the results in Sect. 4.3, we see that the performance of MRSS-ST is unsatisfactory under large scale value. In this work, we propose an Adaptive Midpoint Relay Selection (Adaptive-MRS) scheme that can overcome this shortcoming. Next, we present system design followed by explanation of Adaptive-MRS scheme.

Algorithm 5 Commands performed at Base Station

// Given a source and a destination pair, BS determines candidate relay nodes, maximum allowable transmission power. For a relay selection, a BS also determines whether to probe or not.
Input: s and d
Output: n_{circle}, P_{max} and $Probing$
  Estimate social trust between nodes using call history or available user information
  Determine the emperical values for $Scale$ and $Shape$
  Estimate $Scale_{thres}$ and $Shape_{thres}$
  Receive D2D request from s
  Calculate Midpoint, m = Midpoint(s, d)
  Identify n_{circle} = NodesWithinCircle(m, $radius$)
  if $Scale < Scale_{thres}$ and $Shape < Shape_{thres}$ **then**
    Probe nodes for relay selection i.e. $Probing = 1$
  else
    Select a node nearest to the midpoint as a relay i.e. $Probing = 0$
  end if
  Determine P_{max}
  Send n_{circle}, P_{max} and $Probing$ to s

We assume that the BS knows the social trust nodes have for others. In practise, as implemented in [6, 7, 14], BS obtains this information using call history of users and information collected from online social networks like Facebook and Twitter. With the known values of social trust, the BS fits the social trust values to a Pareto distribution with a particular scale and shape. Then the BS knows the current scale and shape, and uses those values to compare to the scale and shape threshold values denoted as $Scale_{thres}$ and $Shape_{thres}$ respectively. A good estimation techniques can be deployed to determine the threshold values after which the throughput of M-Nearest scheme exceeds MRSS-ST. When the actual scale and shape values exceed the threshold values, a source selects a node that is located nearest to the midpoint and also has a social trust above social threshold (S_{thres}) as a relay. Otherwise, sequential

Algorithm 6 Adaptive-MRS Communication Protocol at Source

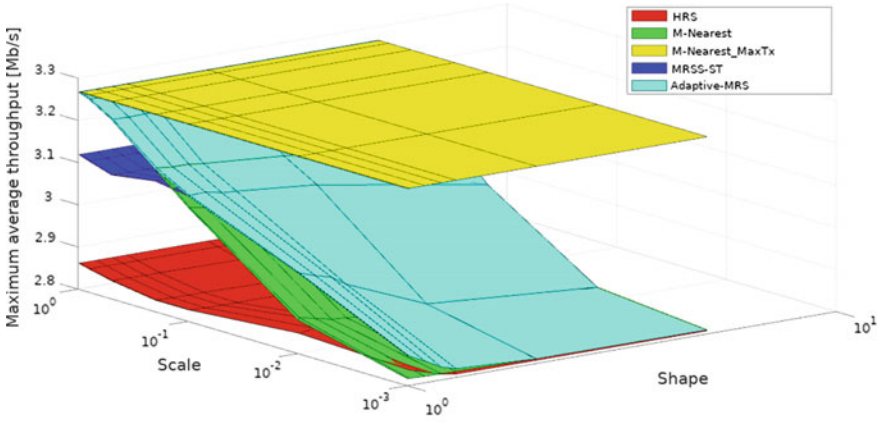
// A source determines whether to use a relay or not. If a relay is to be used, it selects a relay that maximises the throughput of D2D communication.
Input: s, d, n_{circle}, S_{thres}, P_{max}, l, τ and $Probing$
Output: $t_{s,r,d}$

```
if len(n_circle) = 0 then
  Calculate C_{s,d} as in equation 2
  t_{s,r,d} = C_{s,d}
else
  Identify n_{S_thres}
  if Probing = 0 then
    Select a node from n_{S_thres} located nearest to the midpoint as a relay
    Calculate C_{s,r,d} as in equation 3
    t_{s,r,d} = C_{s,r,d}
  else
    if len(n_{S_thres}) = 0 then
      t_{s,r,d} = C_{s,d}
    else
      if len(n_{S_thres}) > l then
        n_{S_thres} = l nodes with max social trust values
        Sequentially probe nodes in n_{S_thres} and d
        Calculate C_{s,r,d} through each of nodes in n_{S_thres} as in equation 3
        Select maximum C_{s,r,d} = max(C_{s,r,d})
        if max(C_{s,r,d}) > C_{s,d} then
          Relay is selected
          t_{s,r,d} = max(C_{s,r,d}) × (T − (τ × n_{S_thres}))
        else
          t_{s,r,d} = C_{s,d} × (T − (τ × n_{S_thres}))
        end if
      end if
    end if
  end if
end if
```

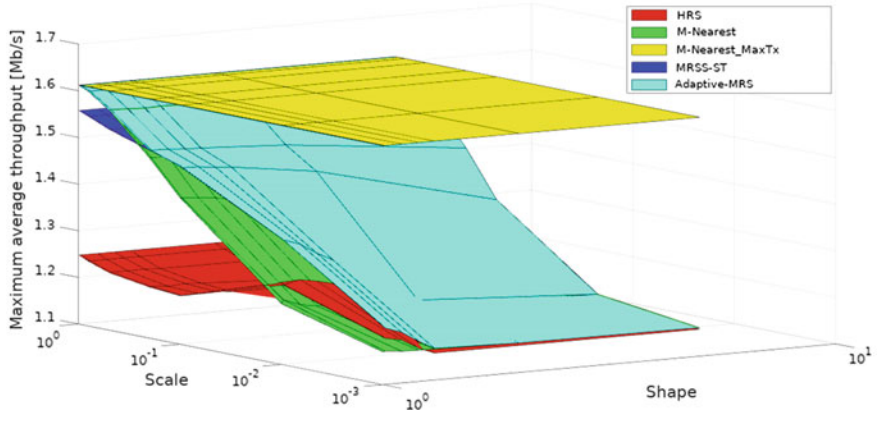
probing of nodes is done for relay selection as shown in Algorithm 5. With the probing of a node, the source knows the signal to noise ratio (SNR) value that incorporates actual transmission power of the node and channel conditions. The actual data rate that can be achieved is calculated by the source as shown in Algorithm 6.

A BS is assumed to adaptively adjust a search radius such that maximum throughput can be achieved. According to Eq. 3, the data rate of D2D communication can be optimised by minimizing the difference between the data rate offered by a link from the source to the relay and that from the relay to the destination. Therefore, a node is selected as a relay which is located within a circular region with a center at the midpoint of the distance between the source and the destination.

A BS identifies the idle nodes that are located within the circular region and sends the list of nodes to the source. The nodes in the list are arranged according to the distance of the node from the midpoint. Among the nodes in the list, at most l nodes are selected to be probed. These nodes are known as candidate relay nodes and have



Network width 100 m



Network width 1000 m

Fig. 5 Maximum of average throughput attained by different relay selection schemes for different node densities

social trust above S_{thres}. The source sequentially probes candidate relay nodes for the relay selection. The source also probes the destination. As a reply to a probe, a candidate relay node sends a SNR value of the received probing packet to the source and also forwards the packet to the destination as in Algorithm 3.

Upon receiving the packets from the candidate relay node and the source, the destination replies back to the source by sending the SNR values as in Algorithm 4. The source decides whether to communicate via a direct link or D2D link after comparing the data rate offered by both the links. The source selects the link offering maximum data rate to maximise throughput as shown in Algorithm 4.

Adaptive-MRS is designed to acheive the performance of MRSS-ST when the social trust among the nodes are low and achieve the performance of M-Nearest when the social trust are high.

After we apply Adaptive-MRS, its performance will be the best of M-nearest and MRSS-ST. Figure 5 shows the comparison of the average throughput of different schemes. This is the ideal situation, where we can estimate the scale and shape with 100% accuracy. However, the result may deviate from this performance because in reality there might be some estimation error. If the estimated scale parameter is not very close to cross-over of MRSS-ST and M-Nearest, we can still expect to get good performance of Adaptive-MRS even with minor estimation error. In the future we will analyse the impact of the estimation error on the performance of Adaptive-MRS. This will allow us to compare Adaptive-MRS to MRSS-ST and other schemes in more realistic scenarios.

6 Conclusion

From the study we can conclude that the average throughput of relay selection schemes is more dependent on the lower limit of the support (scale) as compared to how fast the distribution tail decays (shape). The performance of MRSS-ST is significantly better than performance of HRS in all the scenarios considered. With proper estimation of scale value, proposed Adaptive-MRS scheme can overcome shortcomings of MRSS-ST scheme present when scale value is high.

References

1. Liu, J., Kato, N., Ma, J., Kadowaki, N.: Device-to-device communication in LTE-advanced networks: a survey. IEEE Commun. Surv. Tutor. **17**(4), 1923–1940, (Fourth quarter) (2015)
2. Hasan, M., Hossain, E.: Distributed resource allocation for relay-aided device-to-device communication under channel uncertainties: a stable matching approach. IEEE Trans. Commun. **63**(10), 3882–3897 (2015)
3. Wei, L., Hu, R.Q., Qian, Y., Wu, G.: Energy efficiency and spectrum efficiency of multihop device-to-device communications underlaying cellular networks. IEEE Trans. Veh. Technol. **65**(1), 367–380 (2016)
4. Pan, J.Y., Hsu, M.H.: Relay selection of relay-assisted device-to-device and uplink communication underlying cellular networks. In: Proceedings of the 2017 International Conference on Computing, Networking and Communications (ICNC), pp. 980–985, (2017)
5. Li, C., Jiang, F., Wang, X., Shen, B.: Optimal relay selection based on social threshold for D2D communications underlay cellular networks. In: 2016 8th International Conference on Wireless Communications Signal Processing (WCSP), pp. 1–6, (2016)
6. Pan, X., Wang, H.: On the performance analysis and relay algorithm design in social-aware D2D cooperated communications. In: Proceedings of IEEE 83rd Vehicular Technology Conference, pp. 1–5 (2016)

7. Zhang, M., Chen, X., Zhang, J.: Social-aware relay selection for cooperative networking: an optimal stopping approach. In: Proceedings of the IEEE International Conference on Communications, pp. 2257–2262 (2014)
8. Shrestha Khwakhali, U., Suksompong, P., Gordon, S.: Base station assisted relay selection in device-to-device communications. Int. J. Ad Hoc Ubiquitous Comput [In press]
9. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fMRI: investigating and shaping social mechanisms in the real world. Pervasive Mob. Comput. **7**(6), 643–659 (2011)
10. Li, Y., Wu, T., Hui, P., Jin, D., Chen, S.: Social-aware D2D communications: qualitative insights and quantitative analysis. IEEE Commun. Mag. **52**(6), 150–158 (2014)
11. Chen, X., Proulx, B., Gong, X., Zhang, J.: Exploiting social ties for cooperative D2D communications: a mobile social networking case. IEEE/ACM Trans. Netw. **23**(5), 1471–1484 (2015)
12. Wang, F., Wang, Z., Yang, Z.: Evaluating the influence of social selfishness on cooperative D2D communications. In: Proceedings of the 7th International Workshop on Hot Topics in Planet-scale Mobile Computing and Online Social Networking, HOTPOST '15, pp. 49–54. ACM, New York, USA (2015)
13. Khwakhali, U.S., Gordon, S., Suksompong, P.: Social-aware relay selection scheme for device to device communications in a cooperative cellular network. In: Proceedings of the 2017 International Electrical Engineering Congress, pp. 395–398 (2017)
14. Zhu, X., Du, Q., Ren, P.: Social-aware relay selection for device-to-device underlaying cellular networks. In: 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), pp. 1–5 (2017)
15. Huang, M.L., Coia, V., Brill, P.: A cluster truncated pareto distribution and its applications. In: ISRN Probability and Statistics, vol. 2013, p. 10 (2013)

Collaborative SCM System for Sustainability in the Manufacturing Supply Chain

Donghyuk Jo

Abstract Expansion of the global commercial environment is providing new opportunities and generating crisis for firms at the same time. In the modern dynamic management environment, with rapid changes in technology and diversity of customer needs, firms are facing more difficulties in securing a competitive advantage. Especially for small and medium-sized firms with insufficient resources and competence, it is more difficult to survive and grow due to low competitiveness. Today, small and medium manufacturing firms are required to establish collaborative and mutually beneficial relationships with partner firms in the supply chain to enable them respond quickly to dynamic market changes and gain a competitive advantage in the intensifying global competition. Firms need to overcome the new challenges by creating corporate values and improving performance through collaborative supply chain management (SCM). Therefore, this study has defined the success of export-based small and medium manufacturing firms as export performance, and verified the effects of the collaborative SCM activities of firms and the establishment of the supply chain on the supply chain performance and export performance. The purpose of this study is to identify the importance of collaborative SCM activities to enhance the competitiveness of export-based small and medium manufacturing firms and to suggest strategic directions of SCM.

Keywords Sustainability · SCM system · Supply chain collaboration · Supply chain visibility · Supply chain agility

1 Introduction

Expansion of the global commercial environment is providing new opportunities and generating crisis for firms at the same time. Major firms around the world are overcoming the limitation of their domestic market and are aggressively expanding

D. Jo (✉)
Department of Business Administration, Soongsil University, Seoul, South Korea
e-mail: joe@ssu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_14

to foreign markets for an increase in sales, and export-oriented firms are showing high growth compared to domestic-oriented firms [1].

Steady export growth enhances a firm's global competitiveness and plays a very significant role in the development of the firm and the nation. Export becomes the foundation for a new growth opportunity by allowing domestic firms to overcome the weak points of the domestic market and facilitating the accumulation of capital through quantitative expansion. Also, firms can seek opportunities to emulate and learn advanced technologies in overseas markets and seek improvement of productivity through competition with foreign products. Especially in Korea, a large company-oriented export growth model cannot maintain competitiveness in an ever-changing global environment, and the expectations on the roles of export-oriented small and medium firms are increasing as an alternative. However, the ratio of small and medium firms in the overall export amount in Korea tends to decrease, remaining at a significantly low level compared to major countries like the US and Germany [2].

In the modern dynamic management environment, such as rapid changes in technology and diversity of customer needs, firms are facing more difficulties in securing their competitive advantage. Especially for small and medium firms with insufficient resources and competence, it is more difficult to survive and grow due to low competitiveness. There have been many studies conducted on the overseas expansion and growth of small and medium firms. But studies on the mechanism of achieving continued competitive advantage or export performance by building the capacity, such as seeking new opportunities or securing insufficient resources, against a fast-changing market environment are still extremely inadequate [3, 4].

Today, firms are required to establish collaborative and mutually beneficial relationships with partner firms in the supply chain to respond quickly to dynamic market changes and gain competitive advantages in intensifying global competition and overcome new challenges by creating corporate values and improving performance through collaborative SCM [5–10].

The supply chain collaboration is defined as a relationship in which all firms in the supply chain act to achieve common goals [11]. In other words, the supply chain collaboration is a process in which supply chain firms share information, resources, and risks to achieve common goals to form a long-term partnership. Firms integrate resources and share risks through collaboration to reduce uncertainty and promote the utilization of resources in the supply chain, thereby achieving goals such as enhancement of operational flexibility, cost saving and increased profit [9, 12–17].

The collaborative SCM can visualize the entire processes of supply chain activities and enable the sharing of information between core functions in real-time [18], positively affecting information sharing and collaboration between organizations in the supply chain [13, 19]. In addition, collaborative SCM enables more efficient operation and resource management by facilitating information sharing within the supply chain for demand forecasting and planning [20, 21]. In other words, collaborative SCM increases firms' productivity and efficiency by allowing them to adapt to rapid changes in the marketplace and to shorten the development period of new products [11, 17, 22, 23].

In a management environment where the risks and uncertainties are increasing, it is critical to establish collaborative SCM based on strategic approaches to increase the value of the supply chain to adapt to the rapidly changing environment and thus to secure the supply chain capacities, seeking the growth and long-term survival of firms [24, 25]. The supply chain visibility as a supply chain capability is a key element that strengthens collaboration between supply chain partners, which enables firms to understand better the demand in the supply chain and to achieve better performance in response to customer needs [26]. The supply chain visibility supports the effective use of distributed resources within the supply chain and promotes resource restructuring for more fundamental changes, which in turn affects supply chain performance and ultimately affects the business performance of firms [27]. Also, the supply chain agility is the ability to reconstruct, integrate, and establish the internal and external capacities of a firm to respond to or adapt to a rapidly changing market environment [24]. More specifically, it plays an important role in promoting the coordination and restructuring of resources and enabling the detection of risks and opportunities from the external environment as well as the utilization of the opportunities. This enhances the competitiveness of firms in uncertain market environments [28, 29].

Therefore, this study defined the success of export-based small and medium manufacturing firms as export performance and empirically verified the effects of collaborative SCM activities of firms and the establishment of supply chain capacity through such activities on export performance. The purpose of this study is to identify the importance of collaborative SCM activities to enhance the competitiveness of export-based small and medium manufacturing firms and to establish the strategic directions of SCM.

2 Theoretical Background and Hypotheses

2.1 Supply Chain Collaboration

The collaborative SCM emerges as one of the most important factors that determine the competitiveness of the firm as the risks and uncertainties of the business environment are increasing [30, 31]. The collaborative relationship between firms in the supply chain allows firms to adapt to rapid changes in the market and shorten the development period of new products, thereby enhancing the productivity and efficiency of the firm. In other words, the supply chain collaboration enables the organizations in the supply chain to achieve various goals such as the improvement of organizational capacity, operational efficiency, and customer satisfaction, as well as gain competitive advantages [12, 32, 33].

The supply chain collaboration means that two or more independent firms form long-term relationships and tightly plan and execute supply chain activities toward common goals, thus achieving more benefits compared to when they operate independently [34, 35]. Through collaboration, firms can improve day-to-day operations

and acquire and assimilate the knowledge from their supply chain partners, and the firms in the supply chain can collaborate continuously and effectively with partners through SCM system [21].

The system collaboration as a dimension of supply chain collaboration is an activity to create a mutually compatible communication system to enable forecasting and planning between firms as well as daily electronic transactions and information exchange with supply chain partners [21]. The strategic collaboration is also defined as the combination of forecasted demands by supply chain firms and planned business activities while continued success is being taken into consideration [21, 36]. This strategic collaboration is an essential factor of cooperation and performance between firms in the supply chain [21, 37]. In other words, strategic collaboration first cultivates confidence and commitment of supply chain firms, which in turn strengthens the learning and development of knowledge. Second, developing strategic plans with supply chain partners enables them to understand their roles and act efficiently. Third, strategic collaboration motivates supply chain firms to achieve common goals rather than to seek short-term opportunities individually. Fourth, strategic collaboration helps reduce operating costs [21]. As described above, the supply chain collaboration allows supply chain firms to share long-term partnerships by sharing information, resources, and risks to achieve common goals. More specifically, firms integrate resources and share risks between them through collaboration, which in turn allows them to achieve their common goals such as operational flexibility enhancement, cost reduction and profit maximization [9, 12–17].

The strategic collaboration as supply chain collaboration is based on system collaboration. In other words, a mutually compatible system is required to share demand forecasts and plans between supply chain firms [38]. Therefore, firms need a collaborative information system to establish and institutionalize collaborative relationships with other partner firms. In other words, performing supply chain collaboration activities such as strategic planning and demand forecasting without system collaboration can be a risk [21]. Therefore, this study established hypotheses as follows:

Hypothesis 1 (H1). System collaboration will have a positive effect on strategic collaboration.

The SCM can visualize the entire processes of supply chain activities and enable real-time information sharing between core functions [18], positively affecting information sharing and collaboration between the organizations in the supply chain [13]. Timely and accurate information is essential to the management of products and service flows in the supply chain [27], which is achieved through close interaction between partner firms in the supply chain [9]. In previous studies on the SCM, Fawcett et al. [39] argue that information sharing using information technology in the supply chain enhances the connectivity of the supply chain, thus positively affecting operational performance and customer satisfaction. Brandon-Jones et al. [40] demonstrated that supply chain connectivity as a resource base positively affects information sharing and visibility. Therefore, this study has established hypotheses as follows:

Hypothesis 2 (H2). System collaboration will have a positive effect on supply chain visibility.
Hypothesis 3 (H3). Strategic collaboration will have a positive effect on supply chain visibility.

The collaborative SCM promotes information sharing within the supply chain for demand forecasting and planning, thus enabling operation and resource management in an efficient manner [21]. Swink et al. [37] argue that integration in the supply chain enables the seamless exchange of information between firms, thereby enabling them to predict changes in customer demand, new markets, and technological opportunities, as well as respond faster and better and enhance competitiveness. Also, Koufteros et al. [41] emphasized the role of supply chain collaboration in product development activities, such as product design, process design, and manufacturing activities, and demonstrated the effects of supply chain collaboration on the agility of product development activities. In other words, collaborative SCM increases firms' productivity and efficiency by allowing firms to adapt to rapid changes in the market and shortening the time required to develop new products [12, 17, 21, 32, 33, 37]. Therefore, this study has established hypotheses as follows:

Hypothesis 4 (H4). System collaboration will have a positive effect on supply chain agility.
Hypothesis 5 (H5). Strategic collaboration will have a positive effect on supply chain agility.

2.2 Supply Chain Visibility

It is essential to acquire timely and accurate information for the management of product and service flows in the supply chain [27]. Firms can maximize the efficiency of the firm by checking supply chain planning and execution in real time to identify the entire business flow such as inventory, logistics, and orders. The supply chain visibility is the ability to share and access information across the supply chain [42], and means the access to the information related to the planning and control management of supply chain partners on demand and supply [27]. In other words, information visibility in the supply chain is the ability to access and share information related to the operation and strategies of supply chain partners with the aim to provide partners with the most accurate and latest information on the key activities and processes such as purchasing, manufacturing and distribution [27, 43].

Soejarto [44] argues that identifying and predicting the information trends such as inventory flow, real-time order update, and exception management through process and system integration on the supply chain will facilitate real-time decision making, which will in turn help them find new customers. In addition, Christopher and Lee [45] emphasize that supply chain firms may not know the upstream and downstream operation levels and inventory levels in the supply chain regarding inventory and operational management, which may cause problems with visibility.

As such, ensuring the visibility in the supply chain enhances the transparency of transactions between firms, thereby reducing uncertainty of firms, enhancing trust between firms and enabling the establishment of organic relationships, which in turn allow firms to adapt to changing business environments [26]. Wang and Wei [27] argue that supply chain visibility reduces uncertainty and reinforces confidence, allowing firms to adapt flexibly to changing environments. Li and Zhang [46] say that information sharing and visibility in the supply chain is an important factor in the establishment of an agile supply chain, and that supply chain visibility helps to recognize environmental changes, opportunities and risks in the supply chain. Brandon-Jones et al. [40] suggest that the improved supply chain visibility can lead to improved resilience and robustness by reducing the factors that may dismantle the supply chain. In other words, firms can maximize the efficiency of the firm by identifying the entire supply chain flow such as inventory, logistics, and orders on a real-time basis by analyzing plans and execution in the supply chain [26]. Therefore, this study established hypotheses as follows:

Hypothesis 6 (H6). Supply chain visibility will have a positive effect on supply chain agility.

Also, the supply chain visibility is a key element that strengthens the collaboration between supply chain partners, allowing them to understand better the demands in the supply chain and respond to customer needs [26]. Kulp et al. [47] argue that supply chain visibility improves the efficiency of the supply chain and reduces cycles and inventory shortages. Barratt and Oke [48] maintain that securing visibility through information sharing within the supply chain has a significant effect on operational performance and that operational performance improves through increased visibility. Wei and Wang [49] argue that supply chain visibility reduces supply chain costs and cycles and improves product quality and innovation capacity. Holcomb et al. [50] emphasize the importance of visibility for effective and efficient management of the global supply chain and that supply chain visibility positively affects the competitive position, market share, ROI and customer service level of the firm. In other words, supply chain visibility affects supply chain performance and ultimately affects their business performance by supporting effective use of distributed resources within the supply chain and facilitating resource restructuring for more fundamental changes [49]. Therefore, this study established hypotheses as follows:

Hypothesis 7 (H7). Supply chain visibility will have a positive effect on supply chain performance.
Hypothesis 8 (H8). Supply chain visibility will have a positive effect on export performance.

2.3 Supply Chain Agility

Supply chain agility is the ability of a firm to quickly adjust tactics and operations within the supply chain to respond to or adapt to changes, opportunities, and threats in business environments. In today's uncertain and constantly changing business environment, it receives attention as a dominant competitive lever [25].

The supply chain is a series of related activities involving the design, manufacture, and delivery of products or services within the supply chain. To perform such activities effectively, firms need to collaborate with suppliers in the supply chain to jointly manage market volatility [51]. Therefore, agile supply chain establishment is emerging as one of the most prominent issues of modern SCM and as a strategic approach to increasing supply chain value and adapting to a rapidly changing business environment where uncertainty is intensifying [24, 25, 45].

Christopher [52] defines the supply chain agility as the ability of a firm needed in unexpected and uncertain business environments. He emphasizes that the market sensitivities to detect market changes for agile supply chains, the network between supply chain partners, the virtualization through sharing information between trading partners through information technology utilization, and process integration between supply chain partners are needed for an agile supply chain. Christopher and Lee [45] argue that, to cope with the complexity and uncertainty caused by the changes in the management environment due to globalization, there is a need to strengthen a relationship between the supply chain firms and promptly respond to the changes. This will in turn lead to the reduction of market risks, increased market share, development of new markets and rapid introduction of new products.

Blome et al. [29] define the supply chain agility as the ability of a firm to restructure resources and levels at the supply chain level. Supply chain agility is a critical element of a firm's competitive strategy in an uncertain business environment, and it can provide the firm with a sustainable competitive advantage as the supply chain agility is hard to duplicate. In other words, the supply chain agility is the ability to reconstruct, integrate, and establish the internal and external capacities of a firm to respond to or adapt to a rapidly changing market environment [24]. The agile supply chain capability of a firm can play an important role in promoting coordination and reconstruction of resources. It enables firms to detect risks and opportunities from the external environment and utilize opportunities to improve the competitive advantage of the firm in uncertain market environments [28, 29].

Swafford et al. [24, 53] argue that the firm's agile supply chain capacities, such as short replacement time for materials and services, fast and flexible coordination of production processes, and flexible repositioning of inventory, improve lead times and the coordination of delivery. Tseng and Lin [54] argue that the agile supply chain strategy of a firm contributes to the financial performance of the firm by enabling efficient and rapid response to changing market demands. In other words, agile supply chain capabilities can positively affect the operational and financial performance of a firm by enabling efficient reconstruction and coordination of resources in an

uncertain business environment [29, 55]. Therefore, this study established hypotheses as follows:

Hypothesis 9 (H9). Supply chain agility will have a positive effect on supply chain performance.
Hypothesis 10 (H10). Supply chain agility will have a positive effect on export performance.

2.4 Supply Chain Performance and Export Performance

Firms can enhance efficiencies such as cost savings and product innovation by maintaining collaborative relationships with multiple partner firms in the supply chain [16]. In this way, firms can gain the price competitiveness of the products and services they deliver to customers, improve product and service quality and create new customer values [56]. Increased efficiency of logistics transportation after the manufacture and supply of raw materials and parts reduces unnecessary resource waste and cost burden. Based on these cost advantages, firms can gain price competitiveness in the export market and increase export performance [57]. In other words, improving the efficiency of SCM through collaborative SCM leads to SCM performance such as cost reduction, new product development and improved customer satisfaction, thereby increasing the export performance of manufacturing firms [57]. Therefore, this study established hypotheses as follows:

Hypothesis 11 (H11). Supply chain performance will have a positive effect on export performance.

3 Research Method

3.1 Sample and Data Collection

This study surveyed export-based manufacturers in Korea to validate the proposed research model. The survey was conducted. A total of 345 questionnaires were collected, and 329 cases were selected as valid samples after eliminating missing or inadequate data. The samples of this study are summarized in Table 1.

3.2 Measures

To ensure the content validity of the measurement tool, this study used the measurement items verified in the existing literature by revising and supplementing them

Table 1 Sample characteristics

Category and items		Sample size	Ratio (%)
Operating years	Less than 5 yrs	38	11.6
	5 yrs ~ 10 yrs	76	23.1
	10 yrs ~ 20 yrs	130	39.5
	20 yrs ~ 30 yrs	48	14.6
	More than 30 yrs	37	11.2
Annual sales	Less than $ 10 M	26	7.9
	$ 10 M ~ $ 30 M	41	12.5
	$ 30 M ~ $ 50 M	100	30.4
	$ 50 M ~ $ 100 M	107	32.5
	More than $ 100 M	55	16.7
Industry	Textile/clothing	27	8.2
	Machinery/metal	60	18.2
	Chemicals/energy	54	16.4
	Electric/electronic/communication	119	36.2
	Bio/medical	46	14.0
	Etc.	23	7.0

according to the purpose of this study. This study has developed the measurement items for 'supply chain cooperation', 'supply chain visibility, and 'supply chain agility', 'supply chain performance' and 'export performance' as measurement tools by referring to the study of Kim and Lee [21], Cao and Zhang [33]; the study of Wang and Wei [27] and Williams et al. [58]; the study of Swafford et al. and Gligor et al. [55]; the study of Flynn et al. [9] and Patel et al. [57]; and the study of Katsikeas et al. [59] and Sousa [60], and then measured them by the Likert Seven-point scale (from Not at all to Very Much). The measurement items in this study are summarized in Table 2.

3.3 Analysis Method

For the analysis method and measurement tool of structural equation models, this study analyzed the results and verified the hypothesis using Amos 24.0. For the analysis of the structural equation model, the measurement model was estimated first, and then it was analyzed using the maximum likelihood that is widely used since the two-step approach that estimates the structural model, sample size and the normality assumption were found to be adequate.

Table 2 Confirmatory factor analysis based on reliability

Variable	Measurement items	Factor L.D.	C.R.	Crb. alpha
System collaboration	Accomplishing by electronic communication	0.771	0.905	0.858
	Sharing an information system throughout the collaboration	0.791		
	Working together the information control system	0.799		
	Using the data provided IT tools when making managerial decisions	0.747		
Strategic collaboration	Establishment and prediction of a joint plan is an important/reflected	0.719	0.894	0.802
	Working together strategic issues and policies	0.706		
	Working together planning and executing budget and investment	0.715		
	The performance evaluation and monitoring together.	0.700		
Supply chain visibility	Quick identification of the inventory information	0.761	0.928	0.881
	Quick identification of product information	0.777		
	Quick identification of production plan information	0.862		
	Quick identification of delivery information	0.828		
Supply chain agility	Rapid response to changes in demand/supply	0.753	0.934	0.866
	Quick response to opportunities/threats in the external management environment	0.778		
	Adjustment of short-term production when necessary (due to market demand changes)	0.833		
	Adjustment of product specifications when necessary (due to customer's request)	0.784		
Supply chain performance	Improved inventory turnover of parts and finished products	0.680	0.920	0.834

(continued)

Table 2 (continued)

Variable	Measurement items	Factor L.D.	C.R.	Crb. alpha
	Improved production efficiency	0.794		
	Improved logistics costs	0.803		
	Improved quality satisfaction	0.708		
Export performance	Increased sales in export	0.849	0.912	0.846
	Increased profit rate in export	0.875		
	Satisfaction with export (delivery) performance compared to expectations	0.612		

4 Analysis and Results

4.1 Measurement Model

This study conducted confirmatory factor analysis to ensure the content validity of the measurement tool. For this, χ^2, standard χ^2 (χ^2/df), RMSEA, GFI, TLI, CFI, and IFI were used to check goodness of fit. As a result, initial model did not exceed standard fitness threshold, so modified indices analysis were conducted [61], and measurement items that lowers unidimensionality were deleted EP4. As a result of confirmatory factor analysis of modified measurement model, $\chi^2 = 433.158$ ($P = 0.000$), $\chi^2/df = 2.015$, RMSEA $= 0.056$, GFI $= 0.905$, TLI $= 0.938$, CFI $= 0.948$, IFI $= 0.948$, all indices suggested the measurement model used were fit. After verifying measurement model's fitness, reliability and validity were analyzed. For reliability, construct reliability (C.R.) should appear above 0.7, and average variance extracted (AVE) should be above 0.5. Additionally, for validity, two latent variables' AVE1 and AVE2 should bigger than squared value of its correlation. As a result of analysis, reliability and validity were verified and the detailed results are presented in Tables 2 and 3.

4.2 Structural Model

As measurement model's fitness, and reliability and validity of measurement items were verified, structural model analysis were conducted. As a result of structural model's fitness test, $\chi^2 = 452.174$ ($P = 0.000$), $\chi^2/df = 2.065$ was above threshold 3, and RMSEA $= 0.072$ was below standard of 0.08. Moreover, GFI $= 0.900$, TLI $= 0.935$, CFI $= 0.944$, IFI $= 0.935$ all of indices appeared above recommended value of 0.9 and therefore, the structural model' goodness of fit of the research model was verified.

Table 3 Discriminant validity

Variable	1	2	3	4	5	6
1. Network	**0.681**[a]					
2. Shared vision	0.389	**0.614**[a]				
3. Trust	0.498	0.518	**0.664**[a]			
4. Knowledge sharing	0.448	0.594	0.582	**0.759**[a]		
5. Team efficacy	0.384	0.415	0.462	0.534	**0.795**[a]	
6. R&D performance	0.275	0.364	0.361	0.469	0.476	**0.781**[a]

Note [a]*AVE* (average variance extract)

Table 4 Results of hypotheses tests

Hypothesis	Path	Estimate (β)	C.R. (t)	Supported/ not supported
H1	System collaboration → strategic collaboration	0.604	9.480**	Supported
H2	System collaboration → supply chain visibility	0.278	3.787**	Supported
H3	Strategic collaboration → supply chain visibility	0.532	5.760**	Supported
H4	System collaboration → supply chain agility	0.136	1.941	Not supported
H5	Strategic collaboration → supply chain agility	0.360	3.707**	Supported
H6	Supply chain visibility → supply chain agility	0.280	3.570**	Supported
H7	Supply chain visibility → supply chain performance	0.287	5.014**	Supported
H8	Supply chain visibility → export performance	0.204	2.637**	Supported
H9	Supply chain agility → supply chain performance	0.290	4.761**	Supported
H10	Supply chain agility → export performance	0.375	4.449**	Supported
H11	Supply chain performance → export performance	0.306	2.867**	Supported

*p < 0.05; **p < 0.01; ***p < 0.001

4.3 Hypotheses Tests

After structural model's fitness was confirmed, research hypotheses were tested. As a result, first, for collaborative SCM's structural relationship system collaboration appeared to have an effect on strategic collaboration, $\beta = 0.606$ (C.R. = 9.480, $p = 0.000$), thus, supporting H1. Second, for relationship between collaborative SCM and supply chain visibility, both system collaboration, $\beta = 0.278$ (C.R. = 3.787, $p = 0.000$), and strategic collaboration, $\beta = 0.532$ (C.R. = 5.760, $p = 0.000$) had positive effect on supply chain visibility, therefore, H2 and H3 were supported. Third, for relationship between collaborative SCM and supply chain agility, strategic collaboration had a significant effect on supply chain agility, $\beta = 0.360$ (C.R. = 3.707, $p = 0.000$), while system collaboration had an effect on supply chain agility, $\beta = 0.136$ (C.R. = 1.941, $p = 0.052$), thus H5 was supported while H4 was not supported. Fourth, for relationship between supply chain visibility and supply chain agility, supply chain visibility appeared to have positive effect on supply chain agility, $\beta = 0.280$ (C.R. = 3.570, $p = 0.000$), thus, supporting H6. Fifth, for relationship between supply chain visibility and supply chain performance and export performance, supply chain visibility had a positive effect on supply chain performance, $\beta = 0.287$ (C.R. = 5.014, $p = 0.000$), supply chain visibility had a positive effect on export performance, $\beta = 0.204$ (C.R. = 2.637, $p = 0.008$), therefore, H7 and H8 were supported. Sixth, for relationship between supply chain agility and supply chain performance and export performance, supply chain visibility had a positive effect on supply chain performance, $\beta = 0.290$ (C.R. = 4.761, $p = 0.000$), supply chain visibility had a positive effect on export performance, $\beta = 0.375$ (C.R. = 4.449, $p = 0.000$), therefore, H9 and H10 were supported. Lastly, for relationship between supply chain performance and export performance, supply chain performance had a positive effect on export performance, $\beta = 0.306$ (C.R. = 2.867, $p = 0.000$), thus H11 was supported. The results of hypotheses test are summarized in Table 4.

5 Conclusions

5.1 Summary and Discussion of Results

To identify the success factors of export-based small and medium manufacturing firms, this study suggested supply chain collaboration, supply chain visibility, supply chain agility, and supply chain performance as influence factors of export performance, and empirically verified the path that leads to export performance. Thus, meaningful conclusions have been derived as follows:

First, the system collaboration as a sub-dimension of supply chain collaboration has a positive effect on strategic collaboration. In a supply chain environment, a collaborative information system is required for a firm to establish and institutionalize collaborative relationships with partner firms [21]. Therefore, export-based small and

medium manufacturing firms can continuously and effectively cooperate with partner firms through mutually compatible SCM system in the supply chain.

Second, the system collaboration and strategic collaboration as supply chain collaboration have a positive effect on supply chain visibility. In a supply chain environment, firms need to acquire timely and accurate information to manage product and service flows, and this can be achieved through close interaction between partner firms in the supply chain [27]. Therefore, the collaborative SCM of export-based small and medium manufacturing firms visualizes the entire process of supply chain activities and affects real-time information sharing between core functions positively.

Third, the strategic collaboration as a supply chain collaboration has a positive effect on supply chain agility. In a supply chain environment, collaborative SCM promotes information sharing with partner firms, thus enabling them to predict the changes in customer needs, new markets, and technological opportunities and to respond to the changes faster and better [21]. Therefore, the collaborative SCM of export-based small and medium manufacturing firms has a positive effect on agile SCM by enabling efficient operation and resource management of firms.

Fourth, supply chain visibility has a positive effect on supply chain agility. Supply chain visibility in a supply chain environment reduces uncertainty and enhances confidence to adapt to changing environments' flexibly [26]. Therefore, export-based small and medium manufacturing firms can maximize the efficiency of the firm by securing the visibility of the supply chain and managing the entire flow of the supply chain such as inventory, logistics and orders.

Fifth, supply chain visibility has a positive effect on supply chain performance and export performance. Supply chain visibility is a key element that strengthens collaboration between supply chain partners, allowing them to understand supply chain needs better and respond to customer needs [26]. Therefore, the supply chain visibility of export-based small and medium manufacturing firms has a positive effect on supply chain performance and export performance by promoting effective utilization of resources.

Sixth, supply chain agility has a positive effect on supply chain performance and export performance. Supply chain agility [24] is the ability to reconstruct, integrate and establish firm's internal and external capacities to respond to and adapt to rapidly changing market conditions, playing an important role in improving the competitive advantages of a firm in uncertain market environments [29]. Therefore, the agile supply chain competence in export-based small and medium manufacturing firms has a positive effect on supply chain performance and export performance by enabling efficient restructuring and adjustment of resources in uncertain business environments.

Finally, the supply chain performance has a positive effect on export performance. The manufacturing firms can improve the efficiency by maintaining a collaborative relationship with partner firms in the supply chain, which in turn enables them to secure price competitiveness, creating new customer values [56]. Therefore, the efficient SCM of export-based small and medium manufacturing firms improves supply chain performance, which in turn affects export performance positively.

5.2 Implications and Limitations

This study suggested a path to export performance for export-based small and medium manufacturing firms and demonstrated the theoretical expansion by empirically verifying this. Through literature review, this study defined the success of export-based small and medium firms as export performance and empirically confirmed the effects of the collaborative SCM activities and the establishment of the supply chain capacities on supply chain performance and export performance.

This study has contributed to verifying the importance of the collaborative SCM activities and the establishment of supply chain capacities for enhancing the competitiveness of export-based small and medium manufacturing firms and suggesting the strategic directions of SCM.

This study also allows room for future research. Collaborative SCM could be organized in differently, leading to differences in the success and benefits provides. The collaborative SCM and its relationship to a firm's supply chain capability and supply chain performance would be a useful object of future research.

References

1. Cho, Y.S.: The moderating effects of open innovation on the path toward export performance in the convergence manufacturing SMEs. J. Korea Res. Assoc. Int. Commer. **16**(4), 337–357 (2016)
2. Ha, S.H., Jeong, Y.S., Park, H.H.: A transaction cost approach to analysis on determinants of Korean SMEs' transformation into direct export. Int. Commer. Inf. Rev. **18**(3), 181–201 (2016)
3. Chun, J.I., Yim, H.R.: A study on the effect of firm-specific resources, strategic orientation, and dynamic capabilities on the export performance of Korean exporting SMEs. Korea Trade Rev. **40**(5), 285–313 (2015)
4. Kot, S.: Sustainable supply chain management in small and medium enterprises. Sustainability. **10**(4), 1143 (2018)
5. Wisner, J.D., Tan, K.C.: Supply chain management and its impact on purchasing. J. Supply Chain Manag. **36**, 33–42 (2000). https://doi.org/10.1111/j.1745-493x.2000.tb00084.x
6. Ketchen, D.J., Giunipero, L.C.: The intersection of strategic management and supply chain management. Ind. Mark. Manage. **33**(1), 51–56 (2004)
7. Power, D.: Supply chain management integration and implementation: a literature review. Supply Chain Manag. Int. J. **10**(4), 252–263 (2005)
8. Terpend, R., Tyler, B.B., Krause, D.R., Handfield, R.B.: Buyer–supplier relationships: derived value over two decades. J. Supply Chain Manag. **44**(2), 28–55 (2008)
9. Flynn, B.B., Huo, B., Zhao, X.: The impact of supply chain integration on performance: a contingency and configuration approach. J. Oper Manag. **28**(1), 58–71 (2010)
10. Zhao, L., Huo, B., Sun, L., Zhao, X.: The impact of supply chain risk on supply chain integration and company performance: a global investigation. Supply Chain Manag. Int. J. **18**(2), 115–131 (2013)
11. Mentzer, J.T., Foggin, J.H., Golicic, S.L.: Collaboration: the enablers, impediments, and benefits. Supply chain Manag. Rev. **4**(4), 52–58 (2000)
12. Mentzer, J.T., DeWitt, W., Keebler, J.S., Min, S., Nix, N.W., Smith, C.D., Zacharia, Z.G.: Defining supply chain management. J. Bus. Logist. **22**(2), 1–25 (2001). https://doi.org/10.1002/j.2158-1592.2001.tb00001.x

13. Stank, T.P., Keller, S.B., Daugherty, P.J.: Supply chain collaboration and logistical service performance. J. Bus. Logist. **22**(1), 29–48 (2001). https://doi.org/10.1002/j.2158-1592.2001.tb00158.x
14. Manthou, V., Vlachopoulou, M., Folinas, D.: Virtual e-Chain (VeC) model for supply chain collaboration. Int. J. Prod. Econ. **87**(3), 241–250 (2004). https://doi.org/10.1016/s0925-5273(03)00218-4
15. Sheu, C., Rebecca Yen, H., Chae, B.: Determinants of supplier-retailer collaboration: evidence from an international study. Int. J. Oper. Prod. Manag. **26**(1), 24–49 (2006). https://doi.org/10.1108/01443570610637003
16. Koh, S.C., Demirbag, M., Bayraktar, E., Tatoglu, E., Zaim, S.: The impact of supply chain management practices on performance of SMEs. Ind Manag Data Syst. **107**(1), 103–124 (2007). https://doi.org/10.1108/02635570710719089
17. Wu, L., Chuang, C.H., Hsu, C.H.: Information sharing and collaborative behaviors in enabling supply chain performance: a social exchange perspective. Int. J. Prod. Econ. **148**, 122–132 (2014). https://doi.org/10.1016/j.ijpe.2013.09.016
18. Wong, C.Y., Boon-Itt, S., Wong, C.W.: The contingency effects of environmental uncertainty on the relationship between supply chain integration and operational performance. J. Oper. Manag. **29**(6), 604–615 (2011). https://doi.org/10.1016/j.jom.2011.01.003
19. Liu, C., Huo, B., Liu, S., Zhao, X.: Effect of information sharing and process coordination on logistics outsourcing. Ind Manag Data Syst. **115**(1), 41–63 (2015). https://doi.org/10.1108/imds-08-2014-0233
20. Yu, Z., Yan, H., Edwin Cheng, T.C.: Benefits of information sharing with supply chain partnerships. Ind Manag Data syst. **101**(3), 114–121 (2001). https://doi.org/10.1108/02635570110386625
21. Kim, D., Lee, R.P.: Systems collaboration and strategic collaboration: their impacts on supply chain responsiveness and market performance. Decis. Sci. **41**(4), 955–981 (2010). https://doi.org/10.1111/j.1540-5915.2010.00289.x
22. McIvor, R., McHugh, M.: Partnership sourcing: an organization change management perspective. J. Supply Chain Manag. **36**(2), 12–20 (2000). https://doi.org/10.1111/j.1745-493x.2000.tb00247.x
23. Lee, J.S., Kim, S.K., Lee, S.Y.: Sustainable supply chain capabilities: accumulation, strategic types and performance. Sustainability **8**, 503 (2016). https://doi.org/10.3390/su8060503
24. Swafford, P.M., Ghosh, S., Murthy, N.: The antecedents of supply chain agility of a firm: scale development and model testing. J. Oper. Manag. **24**(2), 170–188 (2006). https://doi.org/10.1016/j.jom.2005.05.002
25. Gligor, D.M., Holcomb, M.C.: Understanding the role of logistics capabilities in achieving supply chain agility: a systematic literature review. Supply Chain Manag. Int. J. **17**(4), 438–453 (2012). https://doi.org/10.1108/13598541211246594
26. Gunasekaran, A., Lai, K.H., Cheng, T.E.: Responsive supply chain: a competitive strategy in a networked economy. Omega **36**(4), 549–564 (2008). https://doi.org/10.1016/j.omega.2006.12.002
27. Wang, E.T., Wei, H.L.: Inter-organizational governance value creation: coordinating for information visibility and flexibility in supply chains. Decis. Sci. **38**(4), 647–674 (2007). https://doi.org/10.1111/j.1540-5915.2007.00173.x
28. Braunscheidel, M.J., Suresh, N.C.: The organizational antecedents of a firm's supply chain agility for risk mitigation and response. J. Oper. Manag. **27**(2), 119–140 (2009)
29. Blome, C., Schoenherr, T., Rexhausen, D.: Antecedents and enablers of supply chain agility and its effect on performance: a dynamic capabilities perspective. Int. J. Prod. Res. **51**(4), 1295–1318 (2013)
30. Lambert, D.M., Cooper, M.C.: Issues in supply chain management. Ind. Mark. Manage. **29**(1), 65–83 (2000)
31. Wang, J., Ran, B.: Sustainable collaborative governance in supply chain. Sustainability **10**, 171 (2018). https://doi.org/10.3390/su10010171

32. Burt, D.N., Dobler, D.W., Starling, S.L.: World Class Supply Chain: the Key to Supply Chain Management. McGraw-Hill, New York (2004). https://doi.org/10.1080/1097198x.2004.10856380
33. Cao, M., Zhang, Q.: Supply chain collaboration: impact on collaborative advantage and firm performance. J. Oper. Manag. **29**(3), 163–180 (2011). https://doi.org/10.1016/j.jom.2010.12.008
34. Simatupang, T.M., Sridharan, R.: The collaboration index: a measure for supply chain collaboration. Int. J. Phys. Distrib. Logist. Manag. **35**(1), 44–62 (2005). https://doi.org/10.1108/09600030510577421
35. Cao, M., Vonderembse, M.A., Zhang, Q., Ragu-Nathan, T.S.: Supply chain collaboration: conceptualisation and instrument development. Int. J. Prod. Res. **48**(22), 6613–6635 (2010)
36. Sanders, N.R., Premus, R.: Modeling the relationship between firm IT capability, collaboration, and performance. J. Bus. Logist. **26**(1), 1–23 (2005). https://doi.org/10.1002/j.2158-1592.2005.tb00192.x
37. Swink, M., Narasimhan, R., Wang, C.: Managing beyond the factory walls: effects of four types of strategic integration on manufacturing plant performance. J. Oper. Manag. **25**(1), 148–164 (2007). https://doi.org/10.1016/j.jom.2006.02.006
38. Ahuja, G.: Collaboration networks, structural holes, and innovation: a longitudinal study. Adm. Sci. Q. **45**(3), 425–455 (2000). https://doi.org/10.2307/2667105
39. Fawcett, S.E., Wallin, C., Allred, C., Fawcett, A.M., Magnan, G.M.: Information technology as an enabler of supply chain collaboration: a dynamic-capabilities perspective. J. Supply Chain Manag. **47**(1), 38–59 (2011). https://doi.org/10.1111/j.1745-493x.2010.03213.x
40. Brandon-Jones, E., Squire, B., Autry, C.W., Petersen, K.J.: A contingent resource-based perspective of supply chain resilience and robustness. J. Supply Chain Manag. **50**(3), 55–73 (2014)
41. Koufteros, X.A., Rawski, G.E., Rupak, R.: Organizational integration for product development: the effects on glitches, on-time execution of engineering change orders, and market success. Decis. Sci. **41**(1), 49–80 (2010). https://doi.org/10.1111/j.1540-5915.2009.00259.x
42. Swaminathan, J.M., Tayur, S.R.: Models for supply chains in E-business. Manag. Sci. **49**(10), 1387–1406 (2003). https://doi.org/10.1287/mnsc.49.10.1387.17309
43. Caridi, M., Moretto, A., Perego, A., Tumino, A.: The benefits of supply chain visibility: a value assessment model. Int. J. Prod. Econ. **151**, 1–19 (2014). https://doi.org/10.1016/j.ijpe.2013.12.025
44. Soejarto, A.: Setting the Stage for Real-Time Enterprise Transformation. Gartner Group (2003)
45. Christopher, M., Lee, H.: Mitigating supply chain risk through improved confidence. Int. J. Phys. Distrib. Logist. Manag. **34**(5), 388–396 (2004). https://doi.org/10.1108/09600030410545436
46. Li, L., Zhang, H.: Confidentiality and information sharing in supply chain coordination. Manag. Sci. **54**(8), 1467–1481 (2008). https://doi.org/10.2139/ssrn.690862
47. Kulp, S.C., Lee, H.L., Ofek, E.: Manufacturer benefits from information integration with retail customers. Manag. Sci. **50**(4), 431–444 (2004). https://doi.org/10.1287/mnsc.1030.0182
48. Barratt, M., Oke, A.: Antecedents of supply chain visibility in retail supply chains: a resource-based theory perspective. J. Oper. Manag. **25**(6), 1217–1233 (2007)
49. Wei, H.L., Wang, E.T.: The strategic value of supply chain visibility: increasing the ability to reconfigure. Euro. J. Inf. Syst. **19**(2), 238–249 (2010). https://doi.org/10.1057/ejis.2010.10
50. Holcomb, M.C., Ponomarov, S.Y., Manrodt, K.B.: The relationship of supply chain visibility to firm performance. Supply Chain Forum Int. J. **12**(2), 32–45 (2011). https://doi.org/10.1080/16258312.2011.11517258
51. Hoek, R., Harrison, A., Christopher, M.: Measuring agile capabilities in the supply chain. Int. J. Oper. Prod. Manag. **21**(1/2), 126–148 (2001). https://doi.org/10.1108/01443570110358495
52. Christopher, M.: The agile supply chain: competing in volatile markets. Ind. Mark. Manag. **29**(1), 37–44 (2000). https://doi.org/10.1016/s0019-8501(99)00110-8
53. Swafford, P.M., Ghosh, S., Murthy, N.: Achieving supply chain agility through IT integration and flexibility. Int. J. Prod. Econ. **116**(2), 288–297 (2008). https://doi.org/10.1016/j.ijpe.2008.09.002

54. Tseng, Y.H., Lin, C.T.: Enhancing enterprise agility by deploying agile drivers, capabilities and providers. Inf. Sci. **181**(17), 3693–3708 (2011). https://doi.org/10.1016/j.ins.2011.04.034
55. Gligor, D.M., Esmark, C.L., Holcomb, M.C.: Performance outcomes of supply chain agility: when should you be agile? J. Oper. Manag. **33**, 71–82 (2015). https://doi.org/10.1016/j.jom.2014.10.008
56. Kushwaha, G.S.: Operational performance through supply chain management practices. Int. J. Bus. Soc. Sci. **3**(2), 222–232 (2012)
57. Patel, P.C., Azadegan, A., Ellram, L.M.: The effects of strategic and structural supply chain orientation on operational and customer-focused performance. Decis. Sci. **44**(4), 713–753 (2013). https://doi.org/10.1111/deci.12034
58. Williams, B.D., Roh, J., Tokar, T., Swink, M.: Leveraging supply chain visibility for responsiveness: the moderating role of internal integration. J. Oper. Manag. **31**(7–8), 543–554 (2013). https://doi.org/10.1016/j.jom.2013.09.003
59. Katsikeas, C.S., Leonidou, L.C., Morgan, N.A.: Firm-level export performance assessment: review, evaluation, and development. J. Acad. Mark. Sci. **28**(4), 493–511 (2000). https://doi.org/10.1108/01443570110358495
60. Sousa, C.M.: Export performance measurement: an evaluation of the empirical research in the literature. Acad. Mark. Sci. Rev. **9**, 1–24 (2004)
61. Hair, J.F., Black, W.C., Babin, B., Anderson, R.E., Tatham, R.L.: Multivariate Data Analysis, 7th edn. Prentice Hall, Englewood Cliffs, New Jersey (1995)

A Study on the Effect of Cultural Capital on the Innovative Behavior

Hye Jung Kim, Jongwoo Park and Myeong Sook Park

Abstract This study provides new perspective by the principle of individual innovation from the perspective of humanitarianism based on the theoretical background that individuals constituting an organization, not the object of the organizational unit, should be the principal agent of corporate innovation to succeed in corporate innovation. In addition, in order to accomplish the purpose of this research, a theoretical model of the 'cycle structure of transformation' was established. It has been confirmed that the cultural capital of an individual embodied by habitus, the class-oriented tendency, argued by Pierre Bourdieu has a positive influence on innovative behavior through mimesis, which instinctively imitates an object for rational self-determination and self-preservation to fulfill needs. Because this study confirms the extent to which individually embodied cultural capital influences innovative behavior and which principles drive individual-level innovative behavior by self-determination and mimesis, the fundamental working principle of enterprise innovation will be clarified. Accordingly, the purpose of this study is to provide theoretical framework and implications needed for the strategic innovation measures, improvement of management techniques, and development of new operational management model in the rapidly changing era. Also, in addition to collective unconscious cultural capital, future studies should research factors that hinder innovation by adding the unconscious perceptual variables such as self-contradiction according to cognitive bias.

Keywords Mimesis · Innovative behavior · Cultural capital

Focused on Mediatory Effects of Self-Determination and Mimsis.

H. J. Kim · J. Park (✉) · M. S. Park
Department of Business Administration, Soongsil University, Seoul, South Korea
e-mail: jongpark7@ssu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_15

1 Introduction

In his book The Singularity is Near in 2006 [1], Ray Kuzweil argued that the humanity improves and evolves in a broader scope with a higher velocity due to the technological singularity, such a mass industrial change has been conceptualized as an industrial revolution by A. J. Toynbee, an economist in the UK, in his book [2] Lectures on the Industrial Revolution of the 18th Century in England. The concept of the industrial revolution of Toynbee was not a particular social phenomenon due to environmental changes over the economy, politics, and society at large, but rather argued that it is a process of progressive and continuative innovation. Accordingly, in today's rapidly changing business environment, companies can derive the innovation of the companies suitable for the rapidly changing business environment by trying to understand the constitutor, the substance of the phenomenon, rather than identifying ideation of social phenomena to establish a response strategy.

These days when the era of the Fourth Industrial Revolution has approached, society has been transformed into the hyper-connected society that anyone can easily connect to the knowledge. That is, without knowledge, the industrial environment has been changed into where any type of economy can be started and developed. Such a trend corresponds with the recognition and argument that the innovative behavior of an innovator who leads the creative disruption argued by Joseph Alois Schumpeter, an economist of the US, in the early days and that it will be the only limiting factor in the future business environment. In the management environment where the type of economy is unclear and limit in, factors are chaotic because the industrial environment changes at the fast speed that prevents the identification of causal relationship after some time, novel and original cases must be derived via behavior-awareness-response. In other words, the production and operation management model of an enterprise should be changed into the method of recognizing the unfolded situation according to behavior and properly responding. Accordingly, in today's chaotic business environment, innovation behavior (IB) at a personal level as an economically limiting factor is gaining attention.

The extent of innovative behavior at a personal level is determined by the extent to which the personal fulfillment of needs is allowed in the social context that an individual is a part of, which also determines what level an individual participates in a task and aims to become proficient in the task [3]. People, however, have internalized recognition and interpretation standard of conformity that accepts daily phenomena and social world of each individual. Such a standard of conformity of an individual is created differently depending on the personal and social position of the individual, the habitus that refers to the collective unconscious will be expressed differently based on the level of experience accumulation of individuals. That is, because the cultural tendency of an individual differs according to the social rank and position [4–8], the voluntary and innovative behavior of an individual will create different results according to the cultural capital embodied based on the social position of the growth process of individuals. In other words, regardless of the personal efforts, voluntary and innovative behavior of an individual will result in different ways according to

the tendency of social ranks. In relation to this, Pierre Bourdieu, a French sociologist, argued that people have differing levels of cultural capital depending on their class, allowing cultural capital act as an important mechanism that drives individual behaviors in his book, <The Distinction> [9].

Based on the theoretical background, in this study, the innovation activities of companies in the real management system are inferred to be greatly influenced by the cultural capital of an individual. That is, it can be understood that imitative mimesis is mobilized based on a social hierarchical structure that creates differently, the embodied cultural capital of an individual influences self-determination to fulfill psychological needs and depending on the level of cultural capital individuals perceive. Thus, individual cultural capital is considered to be a major influence on innovation behavior and therefore applied as an independent variable in this study. Additionally, the assumption of this study is that the result of innovation will differ because the influence on innovation behavior or mimesis according to the personal subjective perception and innovation behavior of self-determination voluntarily determined to fulfill the needs of cultural capital (CC), the collective unconscious created by the social embeddedness that has not bee studied. Based on this assumption, this study self-reported survey was used to measure this and measure personal internal implicit behavior. That is, based on this study, the level of influence of embodied personal cultural capital on innovation behavior and which principle drives innovation behavior at the personal level depending on self-determination and mimesis are verified to identify the fundamental driving principle of company innovation. Therefore, the purpose of this study is to provide the theoretical structure and implications required for developing a new management model, improvement of management method, and strategic innovation strategies of companies that fit the era of rapid changes. In order to achieve this purpose, the following theoretical structure described in Fig. 1 has been set to proceed with the research.

2 Theoretical Background

2.1 Embodied Cultural Capital

Bourdieu present various forms of capital that are not reduced to economic capital in his book <The Forms of Capital> [5] to overcome the class concept in Karl Heinrich Marx's theory of economic determinism. In other words, by extending the concept of capital, which is uniformly defined in existing economic theory, to a comprehensive meaning, he categorized the concept of capital into economic capital, which is the basis of social class structure, and cultural capital, as well as social capital, which are formed based on social relations, and symbolic capital [10]. Cultural capital is the capital needed to reproduce the social class and it is the expression of individuals about themselves in accordance with their socially-recognized status or position, meaning the visual and aesthetic preferences that correspond to the abil-

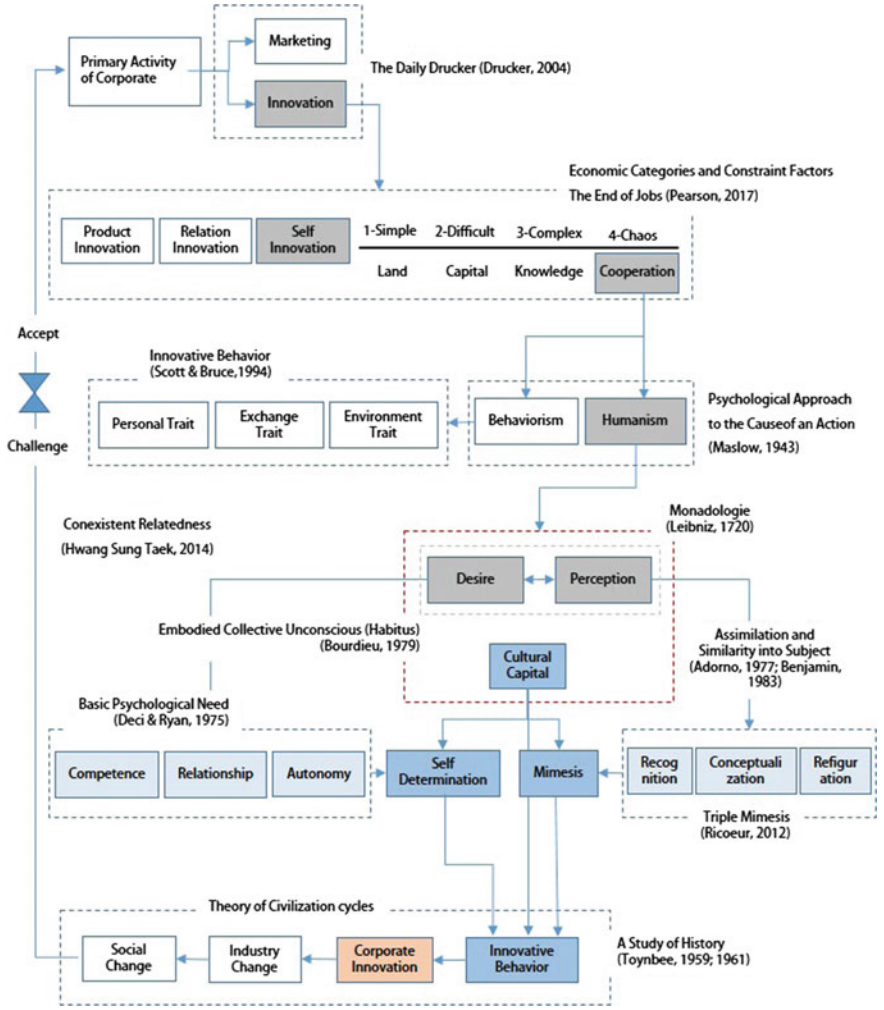


Fig. 1 Cycle structure of transformation

ity to understand and use language assessed by the ruling class of society, personal attitudes, social values, and sensory abilities. That is, cultural capital is total of the habitual preferences and characteristics acquired in the process of socialization and the accumulation and possession of culturally valuable objects as wells the intellectual qualifications including formal qualifications and education. Bourdieu argued the a system of circular relations that integrates social structure and personal behavior by explaining that a layered group is formed according to the ability to enjoy socially valued cultural elements, then, again, within the group, subjective propensity of individuals is created, which causes the personal tendency to lead to behaviors and behaviors again recreate the objective structure of groups.

2.2 Self-Determination Based on Needs

The process of self-selection is critical to the acquisition and accumulation of cultural capital that affects individual social class practices. This self-selection means a rational and reasonable qualification of a person who determines his or her own behavior rather than external pressure or coercive compensation for an act that he or she recognizes and values it as important [11]. Based on the fact that the results differ when individuals have participated in some activities by implicit and external reasons, Edward Deci (1942–) and Richard Ryan (1953–) combined Internal Motivation Theories like Organismic Integration Theory based on Cognitive Evaluation Theory in 1975 to establish the theoretical system of Self-Determination Theory (SDT). This self-determination theory (SDT) is based on the theory of humanism, which recognizes people not only as for survival but as active organisms seeking the optimal way to grow and develop.

2.3 Mimesis Based on Perception

Human beings instinctively try to be like a perceived object at the moment of perceiving the object through experience. In modern times, the concept of mimesis has been expanded its conceptual scope via Walter Benjamin (1892–1940) and Theodor Wiegengrund Adorno (1903–1969) and reconceptualized to refer to a broader meaning that describes the social phenomena of human beings by philosophers such as Paul Ricoeur (1913–2005), who studied phenomenology, and Hans-Georg Gadamer (1900–2002), who criticized modern positivism with the famous expression, "being able to understand is language" and laid the foundations for humanistic hermeneutics. That is, Benjamin broadens the scope of mimesis from an anthropological point of view unlike the general tendency of European scholars who have focused on art since Aristotle [12].

Especially, Benjamin insisted on a primitive mimesis act that mentally reads what was not expressed in language by newly defining mimesis as the ability of nonsensory similarities of becoming mentally intimate with objects that are not mimesis of simple sensory imitation ability [13].

Gadamer [14] positively viewed the meaning of mimesis as an extension of Aristotle and argued that the representation in the work of art is not a mirror that reflects reality, which is a repetition of simply replicating, but rather a re-perception of the essence of the perceived object. This view of Gadamer explains mimesis reproduction as a process of reinterpretation and re-perception of the subject to the essential aspect of reality. This position is continued to narrative hermeneutics of Ricoeur. The theory of mimesis by Ricoeur starts from the perception that all human actions have the nature of time and for this to be truly the time of human beings, it must be narrated and told [15]. In this aspect, his theory of mimesis is referred to as the descriptive mimesis theory. Ricoeur systematizes the mimesis as a theory that explains the for-

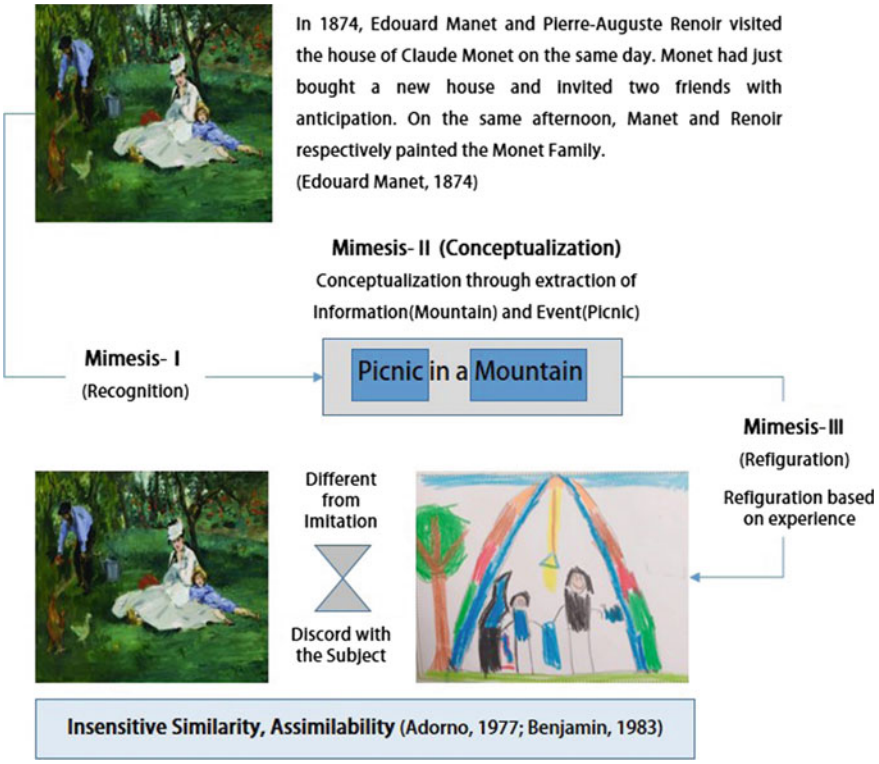


Fig. 2 Explained the 3-layer mimesis

mation process of self-understanding through the narrative act, that is, the process of self-formation, and he called it as the mimesis of 3 layers [16]. Figure 2 explained the 3-layer mimesis.

2.4 Voluntary Innovation Behavior

Innovation at a personal level means acquiring new knowledge and skills and practicing for goals [17]. As such, the concept of innovation behavior is suggested differently by researchers. However, in general, innovation behavior of an individual can be defined as the behavior of voluntarily introducing or developing new ideas on products and services that provide beneficial benefits to organizations and the society [18]. Accordingly, innovation behavior refers to the behavior of the overall process of production operation management of a company because it includes the selection, acceptance, development, implementation, and diffusion of creative ideas.

Table 1 Definition of innovation behavior

Innovation	Starts with finding new and useful innovative ideas [27]
Organization	Relationships with others under a given situation, develop and practice new ideas [28]
Productivity	The process of adopting innovative ideas and transforming them into products, services and processes [29]
Marketing	Promote the ideas and make various efforts to find potential applicants [30]
Profitability	To convert creative ideas to have monetary value [31]

Re-quote from Hwang Sung Taek, 2014

In terms of company management, it refers to every behavior of a businessman who works to successfully practice the creative results [19]. Such a concept of innovation behavior includes the overall process of company management activities that transform the development and acceptance of innovative ideas to substantial monetary value. Thus, innovation behavior that is voluntary and falls outside of the responsibility of a member who constitutes an organization must be evaluated as a strengthening behavior that contributes to the competitiveness of a company [20–23]. Thus, the innovativeness of a company can be perceived as the level of acceptance of innovation and attitude toward practice [24]. That is, it can be identified and measured as the attitude toward innovation, which acknowledges environmental changes and seeks change, and being innovative according to individual traits [25]. Accordingly, this study defines innovation behavior as Table 1. Innovation behavior should be distinguished from creativity. Because creativity is being researched centering on the creative and unique idea, while an individual's innovation behavior is a comprehensive concept that includes the development, acceptance, execution, promotion, diffusion, and conversion of useful ideas that covers the whole of management activities of a corporation. In other words, while creativity is about initiating a creative idea to create, innovation behavior includes the behavior of mimesis that accepts and includes creative ideas of others or developing innovative ideas. In addition, because the concept includes the whole behaviors of utilizing it actively, the practical actions are more emphasized than ideas [26].

3 Research Design

3.1 Research Model

This research infers cultural capital is the fundamental cause of innovation behavior of an individual as each individual possesses in varying degrees according to the social class position. Also, the influence of cultural literacy, cultural activity, and culturally knowledge constituting cultural capital have on innovation behavior at the individual level, and, because cultural capital acquired based on an individual choice

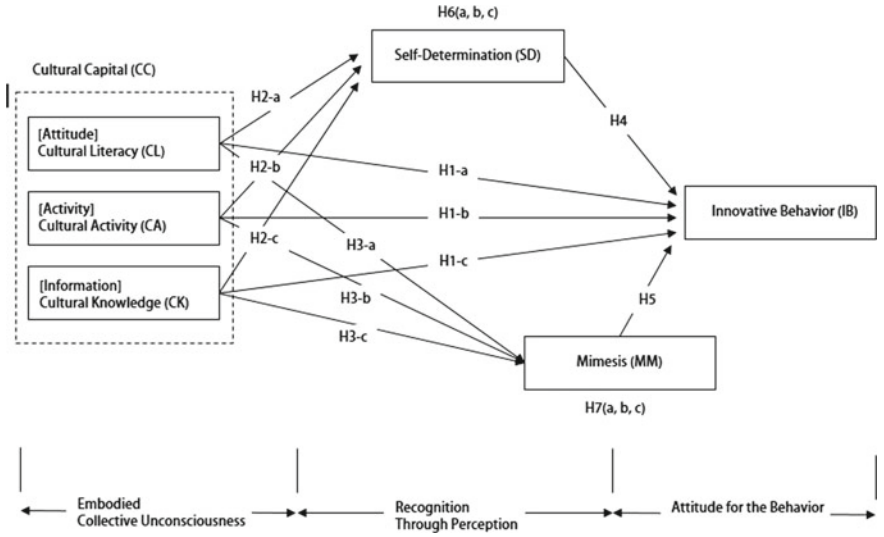


Fig. 3 Research model

fulfills the basic psychological needs of individuals, self-determination and mimesis that instinctively perceives an objective was set as the parameters to empirically examine innovation behavior at the individual level and the following research model was presented as in Fig. 3.

3.2 Operational Definition of Variables

In this study, 7 research hypotheses were established based on theoretical considerations in order to verify the research model empirically. Also, cultural capital, the independent variable, were categorized into cultural literacy, cultural activity, cultural knowledge as conceptualized by Choi and Lee [32] based on DiMaggio [33] to have operational definitions as the degree of interest in specific culture related occupations and activities, in addition to the assessment of cultural refinement, degree of participation in cultural activities, and familiarity and appreciation of specific art genres, Degree of historical understanding.

In addition, for self-determination, the first parameter, the degree to which one believes he can control and regulate his own behavior, which is the operational definition set by Ryan and Deci [34], was applied. For mimesis, the second parameter, the operational definition of the extent to which one tries to resemble or assimilate oneself with a perceived object was created by combining the concepts of Adorno and Horkheimer and Benjamin [35]. Lastly, for the operational definition of innovation behavior, the dependent variable was set by combining the concepts developed by Scott and Bruce [36] and Sullivan and McLean as the degree to which innovative ideas

Table 2 Operational definition of variables

Variable	Lower dimension	Operational definition	Researcher
Cultural capital	Cultural knowledge	The degree of interest in specific culture related occupations and activities, in addition to the assessment of cultural refinement	DiMaggio [33], Choi and Lee [32]
	Cultural activity	Degree of participation in cultural activities	
	Culture understanding	Familiarity and appreciation of specific art genres, Degree of historical understanding	
Self determination		The degree to which on believes he can control and regulate his own behavior	Ryan and Deci [34]
Mimesis		The extent to which on tries to resemble or assimilate oneself with a perceived object	Horkheimer and Adorno [37], [35]
Innovative behavior		The degree to which innovative ideas are developed, accepted, implemented and disseminated	Scott and Bruce [36]

are developed, accepted, implemented and disseminated. The operational definitions are shown in Table 2.

4 Establishment of Hypothesis

4.1 Influence of Cultural Capital on Innovation Behavior

Cultural literacy, cultural activity, cultural knowledge capacity of individuals constituting cultural capital can be understood as a personal tendency that perceives and solves certain problems and the perception ability that constitutes a personal trait which is most influential among antecedents of innovation behavior [29, 38]. Thus, because cultural capital facilitates the subjective perception of the individual. it can

be inferred that it positively affects the individual level of innovation behavior and the following hypotheses were established.

- H1a: Cultural literacy will have a positive influence on innovation behavior.
- H1b: Cultural activity will have a positive influence on innovation behavior.
- H1c: Cultural knowledge will have a positive influence on innovation behavior.

4.2 Influence of Cultural Capital on Self-Determination

Accumulation of cultural literacy, activities, and knowledge of an individual for the acquisition of cultural capital is a socialized human innate behavior to meet rational and efficient basic psychological desire, thus the hypothesis can be set that it will have a positive effect on self-determination.

- H2a: Cultural literacy will have a positive effect on self-determination.
- H2b: Cultural activity will have a positive effect on self-determination.
- H2c: Cultural knowledge will have a positive effect on self-determination.

4.3 Influence of Cultural Capital on Mimesis

At the moment when a human perceives an object, he or she conducts an instinctive mimesis behavior of trying to become similar to and assimilate with the perceived object [39].

Accordingly, the cultural capital of an individual is understood as a result of the instinctive behavior of a human being that aims to become similar and assimilate with a perceived object. Thus, the hypothesis can be set that an individual acquiring homogeneity trying to perceive and become similar to the culture of the upper class and obtaining cultural capital through the methods of cultural literacy, activity, and knowledge will have a positive influence on mimesis.

- H3a: Cultural literacy will have a positive effect on mimesis.
- H3b: Cultural activity will have a positive effect on mimesis.
- H3c: Cultural knowledge will have a positive effect on mimesis.

4.4 The Influence of Self-Determination on Innovation Behavior

People desire to be free to decide what is valuable and important to them, to set goals on their own, and to be the principal agent of innovative behavior [40]. Thus,

self-determination is understood as the rational, reasonable, and fundamental human behavior that is the most efficient way to fulfill their desires in the process of socialization. Thus, a hypothesis can be established that fulfillment of needs such as competence, relationship, and autonomy that constitute an individual's self-determination will have a positive influence on the manifestation and continuation of innovation behavior of an individual.

- H4: Self-determination will have a positive influence on innovation behavior.

4.5 The Influence of Mimesis on Innovation Behavior

Mimesis, which seeks convergence renovation and conducts adaptive convergence in the process of socialization of an individual, is understood as an instinctive behavior of human beings. Accordingly, a hypothesis can be established that mimesis behavior that tries to become similar to and be assimilated with perceived object mimesis will have a positive influence on the manifestation and continuation of innovation behavior of an individual.

- H5: Mimesis will have a positive influence on innovation behavior.

4.6 The Mediating Effect of Self-Determination in the Influence Relationship Between Cultural Capital and Innovation Behavior

- H6a: In the relationship between cultural literacy and innovation behavior, self-determination will play a mediating role.
- H6b: In the relationship between cultural activity and innovation behavior, self-determination will play a mediating role.
- H6c: In the relationship between cultural knowledge and innovation behavior, self-determination will play a mediating role.

4.7 The Mediating Effect of Mimesis in the Influence Relation-Ship Between Cultural Capital and Innovation Behavior

- H7a: In the relationship between cultural literacy and innovation behavior, mimesis will play a mediating role.

- H7b: In the relationship between cultural activity and innovation behavior, mimesis will play a mediating role.
- H7c: In the relationship between cultural knowledge and innovation behavior, mimesis will play a mediating role.

5 Hypothesis Verification

Regression analysis was used to verify the hypothesis that cultural capital influences innovation behavior, self-determination, and mimesis. The verification results for the hypotheses H1, H2, and H3 are as summarized in Table 3.

Hypotheses H1-a, H1-b, and H1-c on the influence of cultural literacy, cultural activity, and cultural knowledge on innovation behavior were selected and to have a positive effect on innovation behavior within the statically significant level.

Table 3 Multiple regression analysis on the relationship between independent and dependent variables

Dependent	Independent	SE	β	t value	Tolerance limit
Innovative	(Invariable)	0.206	–	8.640	–
Behavior	Cultural knowledge	0.060	0.237	3.429**	0.721
	Cultural activity	0.058	0.279	4.238**	0.790
	Cultural understanding	0.018	0.165	2.613*	0.865
R = 0.519, R2 = 0.269, Modified to R2 = 0.259 F = 26.170, p = 0.000, Durbin–Watson = 1.967					
Self	(Invariable)	0.186	–	13.999	–
Determination	Cultural knowledge	0.054	0.275	3.825**	0.721
	Cultural activity	0.052	0.226	3.288**	0.790
	Cultural understanding	0.016	0.057	0.862	0.865
R = 0.451, R2 = 0.204, Modified to R2 = 0.193 F = 18.166, p = 0.000, Durbin–Watson = 1.686					
Mimesis	(Invariable)	0.268	–	8.959	–
	Cultural knowledge	0.077	−0.030	−0.395	0.721
	Cultural activity	0.075	0.211	2.933**	0.790
	Cultural understanding	0.024	0.264	3.836**	0.865
R = 0.358, R2 = 0.128, Modified to R2 = 0.116 F = 10.422, p = 0.000, Durbin–Watson = 1.883					

Standards: *p¡0.05, **p¡0.01

The verification result of the influence of cultural literacy, cultural activity, and cultural knowledge on self-determination showed that Hypotheses H2-a and H2-b on the influence of cultural literacy and cultural activity on self-determination were selected, while Hypothesis H2-c was rejected as the influence of cultural knowledge on self-determination had the t value of 0.862 (p = 0.390). That is, cultural literacy and cultural activity were found to have a positive influence on self-determination within the statistical significance level, while there was no influential relationship between cultural knowledge and self-determination.

The verification result of the influence of cultural literacy, cultural activity, and cultural knowledge on mimesis showed that Hypothesis H3-a was rejected because the influence of cultural literacy on mimesis had the t value of 0.395 (p = 0.693). On the other hand, H3-b and H3-c on the influence of cultural activity and cultural knowledge on mimesis were selected. In other words, although there was no influencing relationship between cultural literacy and mimesis, cultural activity and cultural knowledge were found to have a positive effect on mimesis based on statistical significance.

The results of the hypotheses H4 and H5 that self-determination and mimesis will have a significant effect on the innovation behavior are shown in Table 4.

Hypothesis H4 of the effect of self-determination on innovation behavior was selected and Hypothesis H5 of the effect of mimesis on innovation behavior was selected as well. That is, both self-determination and mimesis were found to influence innovation behavior at statistical significance levels.

Table 4 A simple regression analysis of relationship between independent and dependent variables

Dependent	Independent	SE	β	t value
Innovative	(Invariable)	0.267	–	4.619
Behavior	Self determination	0.068	0.514	8.778**
R = 0.514, R2 = 0.264, Modified to R2 = 0.260 F = 77.051, p = 0.000, Durbin–Watson = 1.770				
Innovative	(Invariable)	0.172	–	9.780
Behavior	Mimesis	0.046	0.605	11.141**
R = 0.605, R2 = 0.366, Modified to R2 = 0.363 F = 124.117, p = 0.000, Durbin–Watson = 2.104				

Standards: *p¡0.05, **p¡0.01

5.1 Verification of the Mediating Effect of Self-Determination in the Relationship Between Cultural Capital and Innovation Behavior

In order to verify the mediating effect of self-determination in the relationship between cultural capital (cultural literacy, cultural activity, cultural knowledge) and innovation behavior, regression analysis was performed at each stage. The analysis results are shown in Table 5.

Hypotheses H6-a and H6-b were selected, and H6-c was rejected. This result shows that because cultural literacy and cultural activity are the results of activities that individuals themselves acquire in socialization and it arises from contextual behaviors, such as innovation behavior, that detect and solve problems, it can be interpreted as affecting the dependent variable without the parameter.

Table 5 Analysis of the effect of self-deterministic mediation on the relationship between cultural capital and innovative behavior

Independent	Parameter	Dependent	Mediated effect verification	Standardized regression coefficient	t value	R^2
Cultural literacy	Self determination	Innovative behavior	Stage 1	0.275	3.825**	0.204
			Stage 2	0.237	3.429**	0.269
			Stage 3 (independent variable)	0.138	2.085**	0.371
			Stage 3 (parameter)	0.357	5.839**	
Cultural activity	Self determination	Innovative behavior	Stage 1	0.226	3.288**	0.204
			Stage 2	0.279	4.238**	0.269
			Stage 3 (independent variable)	0.199	3.161**	0.371
			Stage 3 (parameter)	0.357	5.839**	
Cultural knowledge	Self determination	Innovative behavior	Stage 1	0.057	0.826	0.204
			Stage 2	0.165	2.613*	0.2691
			Stage3 (independent variable)	0.144	2.459*	0.371
			Stage 3 (parameter)	0.357	5.839**	

Standards: *p¡0.05, **p¡0.01

5.2 *Verification of the Mediating Effect of Mimesis in the Relationship Between Cultural Capital and Innovation Behavior*

In order to the mediating effect of mimesis in the influential relationship among cultural literacy, cultural activity, cultural knowledge, and innovation behavior, regression analysis was performed for each step like the previous analysis, and the analysis results are shown in Table 6.

Hypothesis H7-a was rejected, and H7-b and H7-c were selected. The argument of Amabile [29] could be reaffirmed that mimesis, which plays a role of complete mediation between cultural knowledge and innovation behavior is the main factor that acquires the practical knowledge of the habitual knowledge that an individual forms in the process of socialization for a long period of time, while the knowledge and technical ability that an individual voluntarily acquired via mimesis influences the innovation behavior of the individual that discovers and solves a problem.

Table 6 Analysis of mediation effect of mimesis in the relationship between cultural capital and innovative behavior

Independent	Parameter	Dependent	Mediated effect verification	Standardized regression coefficient	t value	R^2
Cultural literacy	Self determination	Innovative behavior	Stage 1	−0.030	−0.395	0.128
			Stage 2	0.237	3.429**	0.269
			Stage 3 (independent variable)	0.252	4.390**	0.498
			Stage 3 (parameter)	0.512	9.817**	
Cultural activity	Self determination	Innovative behavior	Stage 1	0.211	2.933**	0.128
			Stage 2	0.279	4.238**	0.269
			Stage 3 (independent variable)	0.171	3.065**	0.498
			Stage 3 (parameter)	0.512	9.817**	
Cultural knowledge	Self determination	Innovative behavior	Stage 1	0.264	3.836**	0.128
			Stage 2	0.165	2.613*	0.269
			Stage 3 (independent variable)	0.030	0.545	0.498
			Stage 3 (parameter)	0.512	9.817**	

Standards: *p¡0.05, **p¡0.01

5.3 Summary of Hypothesis Verification

The analysis confirmed that cultural literacy, cultural activity, and cultural knowledge, which are sub-factors of cultural capital, have statistically positive effects on innovation behavior. It also found that self-determination plays a mediating role in the relationship between cultural literacy, cultural activity, and innovation behavior among sub-factors of cultural capital, and it could be learned that mimesis plays a mediating role in the relationship between cultural activity, cultural knowledge, and innovation behavior among sub-factors of cultural capital. The results of this hypothesis verification are shown in Table 7.

Table 7 Analysis of mediation effect of mimesis in the relationship between cultural capital and innovative behavior

Hypothesis		Hypothesis contents	Result
H1	H1-a	Cultural literacy will have a positive effect on innovative behavior	Accept
	H1-b	Cultural activity will have a positive effect on innovative behavior	Accept
	H1-c	Cultural knowledge will have a positive effect on innovative behavior	Accept
H2	H2-a	Cultural literacy will have a positive effect	Accept
	H2-b	Cultural activity will have a positive effect on self determination	Accept
	H2-c	Cultural knowledge will have appositive effect on mimesis	Reject
H3	H3-a	Cultural literacy will have a positive effect on mimesis	Reject
	H3-b	Cultural activity will have a positive effect on mimesis	Accept
	H3-c	Cultural knowledge will have a positive effect on mimesis	Accept
H4		Self determination will have a positive effect on innovative behavior	Accept
H5		Mimesis will have a positive effect on innovative behavior	Accept
H6	H6-a	Self determination will play intermediary role in the relationship between cultural literacy and innovative behavior	Accept
	H6-b	Self determination will play intermediary role in the relationship between cultural activity and innovative behavior	Accept
	H6-c	Self-determination will play intermediary role in the relationship between cultural knowledge and innovative behavior	Reject
H7	H7-a	Mimesis will play intermediary role in the relationship between cultural literacy and innovative behavior	Reject
	H7-b	Mimesis will play intermediary role in the relationship between cultural activity and innovative behavior	Accept
	H7-c	Mimesis will play intermediary role in the relationship between cultural knowledge and innovative behavior	Accept

6 Discussion and Implication

In the case of H2-c and H6-c hypotheses, it has been confirmed that cultural knowledge did not positively influence self-determination or the mediating role of self-determination in innovation behavior. As argued by Bourdieu [41], the results reaffirmed the research result of Deci [42] who argued that the acquisition and accumulation of cultural capital is by self-selection of individuals, and the behavior of trying to accumulate cultural knowledge by doing various activities to acquire cultural literacy is a behavior for which an individual motivated and determined, causing the individual to feel the interest and pleasure to manifest and maintain innovation behavior. The quantitative level of cultural knowledge, however, has a distraction effect for an individual in deciding something. Thus, it can be understood that a person is unable to focus on a presented object or an issue. That is because it is difficult for human beings to invest a cognitive effort to process large amounts of information and knowledge, it could be understood that there is a difficulty in cognitively processing information and knowledge in the amount that makes concentration difficult.

Next, in the course of socialization, the mediating effect of mimesis is that individuals acquire practical knowledge that is formed over a long period of time through mimesis.

This learned knowledge becomes the starting point of future actions. In other words, in this study, the argument of Amabile [43] could be reconfirmed that the knowledge and technical abilities an individual acquire through mimesis oneself influences the innovation behavior an individual that finds and solves problems.

Based on the empirical analysis results of this study, the following implications of the academic and practical dimensions could be attained. As implications of the academic dimension, first, because this study identified the principles of innovation behavior by dividing corporate innovations into self-determination by the fulfillment of the psychological needs of people who are the core of management and mimesis instinctively imitating for self-preservation, the reasons for the development and persistence of individual level of innovation behavior were fundamentally determined.

Secondly, it was theoretically confirmed that innovation behavior of an individual is an act that is unconsciously manifested by the cultural capital acquired and accumulated by the individual in the process of socialization and is a result of sustaining and maintaining via self-determination behavior and instinctive imitating behavior. In other words, it could be suggested that corporate innovation is the result of self-determination by individual needs and imitation based on perception.

Third, the antecedents of innovation behavior were expanded from the scope of behaviorism to humanitarianism. That is, individual-level innovation behavior, which has not been identified due to the obsession on the ideological phenomenon until now, has been verified as collective unconscious, and, by expanding the scope of research to self-determination and mimesis, this study has presented and verified a research model that can be expandable and based on a new perspective.

As a practical implication, first, this study has presented the direction of improving the timing and application of the numerous management techniques introduced in terms of the management of corporations.

Next, it will be possible to secure and maintain a specialized market in terms of business strategy by developing differentiated products and services based on the results of this study.

In this study, in order to understand corporate innovation to adapt to social change and to measure the subjective cultural capital of individuals, an empirical analysis was conducted after collecting samples centering on departments of universities and graduate students related to office work related to the cultural field. Thus, the generalization of the research results is limited due to the demographic specificity of the samples and, thus, there are limits to applying unilaterally to innovation activities of various types of companies.

7 Limitations and Future Research

In addition to the collective unconscious cultural capital proposed in this study, future research will be able to provide a theoretical basis for understanding the individual level of innovation behavior by people's contradictory behavior, such as self-contradiction according to cognitive bias, from various levels if the research is conducted by adding the perception variable about various unconsciousness. In future research, it is necessary to develop and demonstrate research models that can be applied to various situations by merging behavioral factors and humanistic factors that explain individual-level innovation behavior because the research model presented in this study alone cannot explain all the innovation behavior caused by an individual's unconsciousness.

References

1. Kurzweil, R.: The Singularity Is Near: When Humans Transcend Biology (2006)
2. Toynbee, A.J.: Lectures on the Industrial Revolution of the 18th Century in England, London, NY, vol. 1890, pp. 31–32 (1884)
3. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp. Educ. Psychol. **25**, 54–67 (2000)
4. Bernstein, B.: Class, Codes and Control Volume 3: Towards a Theory of Educational Transmissions. Routledge and Kegan Paul, London (1975)
5. Bourdieu, P.: The forms of capital. In: Richardson, J.G. (ed.) Handbook of theory and research for the sociology of education, p. 241. Greenwood Press, Westport (1986)
6. Marshall, A.W., Olkin, I.: A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families (1997)
7. Katz-Gerro, T.: Cultural consumption research: review of methodology, theory, and consequence. Int. Rev. Sociol. **14**(1), 11–29 (2004)

8. Tomlinson, C.: Fulfilling the Promise of the Differentiated Classroom: Strategies and Tools for Responsive Teaching. Association for Supervision and Curriculum Development, Alexandria (2003)
9. Bourdieu, P.: La distinction: Critique sociale du jugement (Paperback) (A Social Critique of the Judgement of Taste, trans. Richard Nice, 1984). Harvard University Press, Cambridge (1970)
10. Cho, K.I.: Mode of leisure consumption and cultural capital: Bourdieu's cultural theory. Tour. Res. **30**(1), 379–401 (2006)
11. Deci, E.L., Ryan, R.M.: The general causality orientations scale: self-determination in personality. J. Res. Personal. **19**, 109134 (1985)
12. Cheong, S.H.: The meaning of mimesis in Adorno's thought. J. New Korean Philos. Assoc. **70**, 423–450 (2012)
13. Benjamin, L.S.: A clinician-friendly version of the interpersonal circumflex: structural analysis of social behavior (SASB). J. Pers. Assess. **66**, 248266 (1996)
14. Gadamer, H.-G.: Philosophical Hermeneutics (Trans. David E. Linge (ed.)). University of California Press (1977)
15. Kim, D.Y.: Reinterpretation des notions aristoteliciennnes (mimesis, muthos, catharsis) et reflexion sur la theorie narrative autour du Temps et recit de Paul Ricoeur. J. Humanit. **32**, 151168 (1997)
16. Lee, J.H.: A Study of paul ricoeur's narrative hermeneutics based on the process of self-becoming. Study Moral Educ. **20**(2), 49–75 (2008)
17. West, M.A.: Innovation among health care professionals. Soc. Behav. **4**, 173–184 (1989)
18. West, M.A., Farr, J.L.: Innovation at work. In West, M.A., Farr, J.L. (eds.) Innovation and Creativity at Work, pp. 1–13 (1990)
19. Oldham, G.R., Cummings, A.: Enhancing creativity: managing work for the high potential employee. Calif. Manag. Rev. **40**(1), 22–38 (1997)
20. Porter, M.E.: The Competitive Advantage: Creating and Sustaining Superior Performance. Free Press, NY (1985). (Republished with a new introduction, 1998.)
21. Ancona, D., Caldwell, D.: Management issues facing new product teams in high technology companies. In Lewin, D., Lipsky, D., Sokel, D. (eds.) Advances in Industrial and Labor Relations, vol. 4, pp. 191–221. JAI Press, Greenwich (1987)
22. Shalley, C.E.: Effects of coaction, expected evaluation, and goal setting on creativity and productivity. Acad. Manag. J. **38**, 483–503 (1995)
23. McCarthy, A.M., Schoenecker, T.S.: Commitment to innovation: the impact of top management team characteristics. R&D Manag **29**, 199–216 (1999)
24. Damanpour, F.: Organizational innovation: a meta-analysis of effects of determinants and moderators. Acad. Manag. J. **34**, 555–590 (1991)
25. Ettlie, E., O'Keefe, R.D.: Innovative attitudes, values, and intentions in organizations. J. Manag. Stud. **2**, 163–182 (1982)
26. Katz, D., Kahn, R.L.: The social psychology of organizations. Wiley, New York (1978)
27. Kanter, R.M.: Three tiers for innovation research. Commun. Res. **155**(5), 509–523 (1988)
28. Van de Ven, A.H.: Central problems in the management of innovation. Manage. Sci. **32**(5), 590–607 (1986)
29. Amabile, T.M.: A model of creativity and innovation in organizations. In Staw, B.M., Cummings, L.L.: Research in Organization Behavior, pp. 187–209. JAI Press, Greenwich (1988)
30. Galbraith, R.C.: Just one look was all it took: reply to Berbaum, Markus, and Zajonc (1982)
31. Rothwell, R.: The relationship between technical change and economic performance in mechanical engineering: some evidence. In: Industrial innovation, pp. 36–59. Palgrave Macmillan, London (1979)
32. Choi, S., Lee, M.J.: Conceptualization of an Index of cultural capital and its measurement with a focus on diMaggios framework. Korean J. Sociol. **47**(2), 31–60 (2013)
33. DiMaggio, P.: Cultural capital and school success: the impact of status culture participation on the grades of U.S. high school students. Am. Sociol. Rev. **47**(2), 189–201 (1982)
34. Ryan, R.M., Deci, E.L.: Overview of Self-Determination Theory: An Organismic Dialectical Perspective. University of Rochester Press, Rochester (2002)

35. Benjamin, L.S.: Inclusive or contested? Conceptualizing a globalized Bangalore. In: Mahadevia, D. (ed.) Inside the Transforming Urban Asia Policies, Processes, and Public Action. Concept, New Delhi (2008)
36. Scott, S.G., Bruce, R.A.: Determinants of innovation behavior: a path model of individual innovation in the workplace. Acad. Manag. J. **37**(3), 580–607 (1994)
37. Horkheimer, M., Adorno, T.: La industria cultural. Industria cultural y sociedad de masas, Caracas, Monte Avila (1969)
38. Howell, M., Higgens, C.A.: Champions of technological innovation. Adm. Sci. Q. **35**, 317–341 (1990)
39. Benjamin, L.S.: Interpersonal Diagnosis and Treatment of Personality Disorders, 2nd edn. Guilford Press, New York (1996)
40. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. **55**, 68–78 (2000)
41. Bourdieu, P.: La distinction: Critique social du jugement. Editions du Minuit, Paris (1979)
42. Deci, E.L.: The Psychology of Self-Determination. Lexiton Books, Lexiton (1980)
43. Amabile, T.M.: How to kill creativity. Harv. Bus. Rev. **76**, 77–87 (1998)

Evaluation of Technology Transfer Performance in Technology-Based Firms

Donghyuk Jo and Jongwoo Park

Abstract This study aims to determine the factors effecting the performance of the firms with transferred public technology and suggests implications thereof. To that end, this study empirically analyzed the effects of transferred technology value (technology transaction amount, technology readiness level) and absorptive capacity (potential absorption capability, feasible absorption capacity) on technology transfer performance. The results of this study show that technologies with a higher technology readiness level have a positive effect on the technology transfer performance of firms. The study also shows that absorptive capacity has a moderating effect on the relationship between technology value and the technology transfer performance of firms, compared to technology with a higher transaction amount. This study is meaningful in that it suggests strategic directions for the successful implementation of technology transfer by confirming the importance of technology value and the necessity of strengthening absorption capacity.

Keywords Technology-based firms · Technology transfer · Technology value · Technology readiness level · Absorptive capacity

1 Introduction

The problem of how to efficiently and effectively create, dissipate, and utilize new knowledge and technologies within the structure of today's knowledge-based economy is becoming an important issue for the competitiveness of individual economic entities, as well as national economic growth [1]. Rapid change and development of technology demands an active and organic linkage among technological innovation actors; also, maximizing the benefits created by the activities of technological inno-

D. Jo (✉) · J. Park
Department of Business Administration, Soongsil University, Seoul, South Korea
e-mail: joe@ssu.ac.kr

J. Park
e-mail: jongpark7@ssu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4_16

vation actors is emphasized as an important policy paradigm to strengthen national competitiveness [2].

In other words, it is necessary to facilitate the flow of technical knowledge by activating technology transfer between among technological innovation actors so that a they can supply excellent technologies can be supplied and Certain conditions need to be met for the successful introduction of external technologies. Therefore, effective government policies and support are needed to ensure that the R&D achievements of public research institutes, which occupy 80% or more of domestic government R&D budgets, are faithfully transferred to private firms and used economically [3].

Bozeman and Crow [4] defined technology transfer as a process of the movement of physical design, processes, know-how, and information from one organization to another. According to the "Act on the Promotion of Technology Transfer and Commercialization" in Korea, the transfer of technology means the transfer of technology from the technology owner (including the person who has the right to sell) to other parties through transfer, licensing, technology grants, technical guidance, joint research, joint venture, merger, or acquisition. In technology transfer, technology is considered to include know-how, in addition to patents and utility models in the broad sense of intellectual property. In particular, from the perspective of a company, technology also includes the information necessary for production and sales of products or services [3, 5].

In Korea, beginning with the amendment of the "Act on the Promotion of Technology Transfer and Commercialization" in 2001, the government has made noticeable improvements by actively promoting the transfer of technology developed through public R&D activities to private firms, and the commercialization of such technologies. According to the KIAT report [6], the number of domestic technology transfers steadily increased, from 4,259 in 2010 to 8,524 in 2014, and technology licensing fee income also steadily increased, from KRW 124.5bil in 2010 to KRW140.3bil in 2014, whereas the technology transaction amount per case decreased from KRW29.24mil in 2010 to KRW16.46 mil in 2014. In addition, the ratio of transferred technology to technology transfer performance of firm decreased from 14.1% in 2010 to 12.4% in 2014.

The technology transaction amount is calculated by valuing the present value of all the benefits expected by such technologies, implying that the higher the technology transaction amount, the higher the technology transfer performance of the firms [7]. However, as mentioned earlier, the question arises as to whether the technology transaction amount has a positive effect on the performance of firms. Existing studies of technology transfer have focused on the types of technologies that can be transferred easily, the factors affecting technology transfer, the methods of calculating technology value, how to sell the technology well, and whether to receive appropriate price of technology, from the viewpoint of technology sellers. As technology commercialization has entered into full swing, and sufficient cases of technology transfer have accumulated in Korea, it is now necessary to study technology value from the viewpoint of technology buyers.

Therefore, the purpose of this study is to determine the factors affecting the technology transfer performance of firms with transferred technology, and to provide

implications for them through the cases of public technology transfer. The information on feasible technology transfer is a topic that has not been studied much, because it is very difficult to collect suitable cases since the information is directly related to the business directions of firms. To this end, this study will suggest the technology value as technology readiness level, which indicates the technology transaction amount in which the technology is converted into a monetary value, and the level of technology completeness. It will empirically verify the effects of technology value on the technology transfer performance of firms, and the moderating effects of the technology absorptive capacity of firms on technology value and technology transfer performance. The study will also aim to confirm the importance of technology value and the necessity of strengthening technology absorptive capacity in order to improve technology transfer performance, thereby suggesting strategic directions for the successful implementation of technology transfer.

2 Theoretical Background and Hypotheses

2.1 Definition and Importance of Technology Value

Technology value varies, based on the position and purpose of the evaluators. Capon and Glazer [5] and Boer [8] defined technology as an intellectual asset based on commercial value. The commercial value referred to here is related to fair market or monetary value. In terms of the management environment, technology is an essential catalyst for creating wealth, and technology plays a very important role in creating corporate value, in conjunction with other firm resources [9]. Firms are paying more attention to technology valuation and packaging as a way to enhance competitiveness [10].

The monetary technology valuation amount is a series of activities done to calculate the market value of technology by clarifying the value of the targeted technologies and analysing the opportunity and risk factors for economic efficiency, rights, substitution, and other factors comprehensively. The technology valuation means expressing the present value of all future technology benefits that are expected to occur. Smith and Parr [11] argued that "It is not possible to create economic benefits with technology alone. The economic benefits can be created by using the technology in conjunction with the business ability, financial ability, and other tangible assets of the firms". They defined the value of technology as part of the economic, technological, and strategic value gained by holding technology, and technology valuation as measuring the value attributable to technology from all economic sources [11]. The technology valuation is an operation to calculate the monetary value of technology, and the technology value amount is used as the basic data for calculating the technology transaction amount, so it is critical to technology transactions.

Furthermore, firms with transferred technology should achieve management performance improvement through the production and sale of the products created by

Table 1 Technology readiness levels (TRL)

TRL	Description
TRL 1	Basic principles/experiments
TRL 2	Technology concept formulated for application purpose
TRL 3	Proof of basic concept in Laboratory
TRL 4	System validation in laboratory
TRL 5	Prototype development and validation (scale: one to several numbers)
TRL 6	Prototype development and validation (scale: beyond percent of mass-production)
TRL 7	Reliability evaluation
TRL 8	Certification and standardization
TRL 9	Commercialization

Note KEIT [20], Ahn et al. [15]

the technologies. The transfer of technology involves unexpected costs and risks because of the large technology gap between the technologies produced and supplied by public research institutions and the technologies required in the market [12]. Therefore, the technology development state is an important technology transaction stage, and this stage is referred to as Technology Readiness Level (TRL). TRL, in nine levels from TRL 1 to TRL 9, was developed and suggested by NASA's Sadin et al. [13] and Mankins [14] to identify the progress of R&D in the aerospace sector [15]. In Korea, Hwang et al. [16] reported that the accurate perception of technology development stages could prevent business delay and budget increases, and enhance business evaluation scores. In the United States, it is also used for national research and development in the defense and energy fields [17, 18].

In Korea, it was used for construction and transportation technology and industrial source strategy technology [19, 20]. Ettlie [21], Lee [22], and Park et al. [7] said that that the higher the TRL of application/development technology, the greater the possibility of commercialization. TRL has been translated into technology maturity and technology readiness in Korea. Recently, technology buyers have used it in order to understand the completeness of technology [15]. The definitions of TRL are listed in Table 1.

2.2 Technology Value and Technology Transfer Performance

In today's technology-intensive industries, technology is changing at a rapid pace, which makes it difficult for firms to maintain profitability and competitive advantage with existing products alone [23]. Therefore, as change in technological innovation accelerates, there is growing interest in acquiring technology from outside, and technology transfer is increasing rapidly in Korea, according to the technology transaction statistics [6].

Existing studies have focused on the relationship between patents, which is a representative measurement index of technology, and technology transfer performance [24, 25]. According to a KIAT report [6], among the number of technology transfers through domestic technology sales in 2014, patents accounted for 80% of all sales, indicating that firms have their own technologies even if they are entering into patent license agreements. Therefore, this study has reviewed the existing studies conducted on the effects of patent license on financial performance.

Ernst [24] reported that the number of patent applications contributes to a company's sales growth performance. He used sales growth rate, sales growth rate per employee, relative sales growth rate, and relative sales growth rate per employee, as a sales performance index. Yam et al. [26] used average sales growth rate and operating income growth over the previous three years. Branch [27] studied whether there was a time lag between expenditure, patent acquisition, and profitability. Considering the fact that firms' profits have a significant effect on expenditure, and these two variables are simultaneously influenced by third-party external factors, analysed the effect of patent holdings on a company's accounting profit margin. As a result, he concluded that activities that have been replaced by patents affect the increase profitability of firms, with several years' lag.

As indicated in the results of previous studies, acquiring technologies positively affects the performance of firms. However, most studies focused on whether the company's patent acquisition affected the company's business performance from the viewpoint of securing the technology, and there are not many studies conducted on technology import from outside the firm. Therefore, technology is used not only to integrate protected inventions into new products, new processes, and new service development, but also to license it to other firms and raise funds through negotiations [28]. In other words, if technology is acquired by paying the technology transaction amount, it should be possible to realize profit through various methods, such as capital formation, licensing, and securing liquidity using old technology. The fact that the technology transaction price is high indicates that the value of the technology is high, and the possibility of realizing future profit is also high. Furthermore, a higher technology readiness level can reduce costs and shorten the time needed to achieve commercialization due to the elimination of technological gaps, which in turn shortens the time need to attain profits. Therefore, this study establishes the following hypotheses:

Hypothesis 1 (H1) Technology transaction amounts will affect Technology transfer performance.
Hypothesis 2 (H2) Technology readiness levels will have a positive effect on Technology transfer performance.

2.3 Absorption Capacity

Firms conduct activities jointly with external organizations to acquire new technological knowledge. For example, firms can create results by commercializing technology acquired through cooperation with public institutions, such as university research institutes or government-backed research institutes. However, not all of the technical knowledge transferred through this R&D cooperation is utilized to show performance. In other words, even if the same technical knowledge is transferred, an identical effect does not occur, because the ability to absorb transferred technology varies from company to company [29].

Absorptive capacity is the ability of a company to acquire technical knowledge and to utilize it efficiently. Absorption capability allows the company to accept new technical knowledge from the outside and convert it into new knowledge to create the business performance of the company [30].

This absorptive capacity is divided into potential absorptive capacity and realized absorptive capacity [31]. Potential absorptive capacity is the ability of a company to identify, understand, and assimilate new external technological knowledge. In addition, realized absorptive capacity refers to the ability of firms to transform and exploit technological knowledge acquired externally to create profits by utilizing technology acquired through potential absorptive capacity [2, 32]. As such, absorptive capacity can be a source of company competitive advantage [33], and firms with high absorptive capacity can absorb external knowledge to achieve efficient results [34–36].

According to Zahra and George [34], firms with high levels of absorption can use their first mover advantage to respond to customers quickly and avoid lockout effects or competency traps, thereby creating excellent results. Therefore, this study establishes the following hypotheses:

Hypothesis 3 (H3) Potential absorptive capacity will have a moderating effect in the relationship between technology value and technology transfer performance.
Hypothesis 4 (H4) Realized absorptive capacity will have a moderating effect in the relationship between technology value and technology transfer performance.

Based on the above hypotheses, study model in this study has been suggested as shown in Fig. 1.

3 Research Method

3.1 Samples and Data Collection

In this study, to validate the established study model, the financial information of firms in relation to 415 technology transfer cases in which the technology was transferred from domestic universities and public research institutes to SMEs from 2008 to 2013,

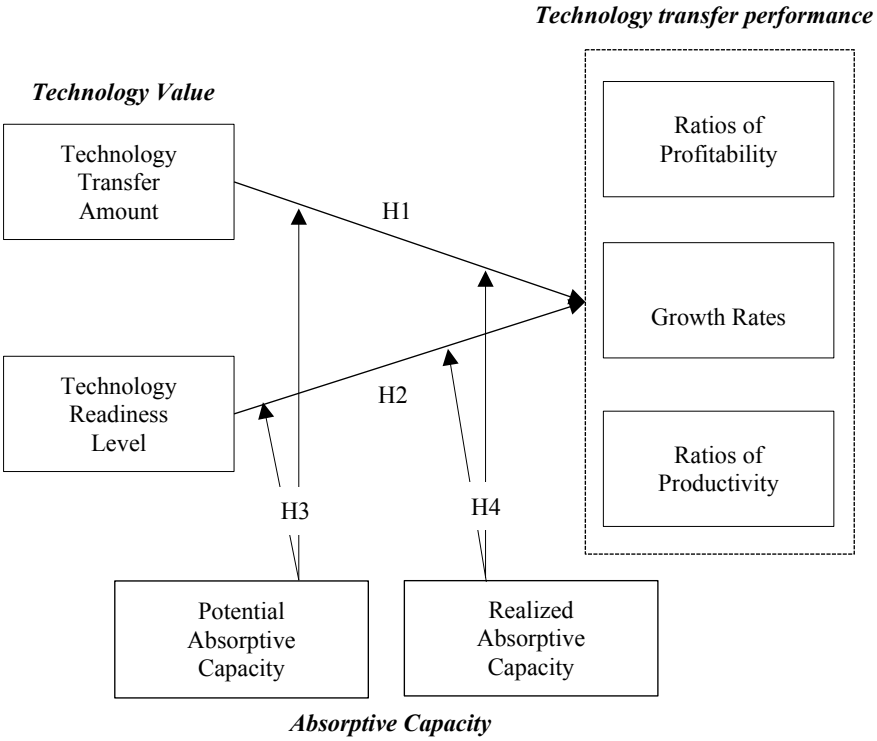


Fig. 1 Research model

was collected using the corporate management information system (www.cretop.com). After removing 96 cases due to closed businesses or insufficient financial data at the time of this study, 319 cases were selected as valid samples. The samples used in this study are listed in Table 2.

Table 2 Sample characteristics

Category	Minimum	Maximum	Mean	Standard deviation
Operating years	5	60	16.99	10.135
Number of employees	3	1278	95.05	136.453
Technology readiness level (1–9)	3	7	4.65	0.741
Technology transfer amount (KW)	1,000,000	220,000,000	39,852,132	41,464,631

3.2 Measures

The technology value factor was measured after dividing it into its sub-dimensions: the technology transfer amount and the technology readiness level. In general, the technology transfer amount is divided into lump sum payment, initial payment, and running royalty [6], and running royalty is difficult to quantify because it is paid when the transferred technology is successfully commercialized. Therefore, in this study, the lump sum payment and initial payment, which are fixed amounts of royalty, were measured using the amount data before the technology transfer. In addition, the degree of technological completeness was measured using the value of technology development stage measured at the time of technology transfer [12, 14, 16].

The technology transfer performance factor was measured after dividing it into its sub-dimensions: the index of the ratios of profitability, the index of growth rates, and the index of ratios of productivity. In addition, since the technology transfer performance should be measured after the technology transfer, the mean value was measured for the three years after the technology transfer. For the growth rate of sales as ratios of profitability index, sales (t) versus sales (t – 1) was measured. For the net income to sales as the index of growth rates, this year's net profit (t) versus previous year's net profit (t – 1) was measured. For the growth rate of sales per capita as the index of the ratios of productivity, the growth rate of sales versus the number of employees was measured. The absorptive capacity of the firms was measured after diving it into its sub-dimensions: the potential absorptive capacity and the realized absorptive capacity.

For the potential absorptive capacity and the realized absorptive capacity, the R&D intensity and the ratios of operating revenue, were set as the measurement index, respectively. In addition, since the absorption capacity is the ability possessed when a new technology is introduced, the year immediately before the technology transfer was established as the analysis time point. In addition, the size and the duration of the firm were set as control variables in this study. In general, the size and duration of firms can have a potential impact on the firms' ability to absorb external resources and develop new technologies and products [34, 37, 38].

The number of employees was set as a measurement index for the size of the firm. In addition, for firm duration, the period from the year of founding until the date of the study's data collection was set as the measurement index. The variables used in this study are listed in Table 3.

4 Analysis and Results

4.1 Main Effect Model

As a result of analysing the effect of technology transaction amount on technology transfer performance, we found that technology transaction amount has a significant

Table 3 Relationship between technology transfer amount and technology transfer performance

Independent variable	Dependent variable	β	t	–
Technology transfer amount	Ratios of profitability	−0.146	−2.636**	$R^2 = 0.021$, F = 6.948**
	Growth rates	0.080	1.421	$R^2 = 0.006$, F = 2.020
	Ratios of productivity	−0.151	−2.718**	$R^2 = 0.023$, F = 7.385**

Note *p < 0.05, **p < 0.01, *** < 0.001

Table 4 Relationship between technology readiness level and technology transfer performance

Independent variable	Dependent variable	β	t	–
Technology readiness level	Ratios of profitability	0.205	3.732***	$R^2 = 0.042$, F = 13.924***
	Growth rates	0.443	8.787***	$R^2 = 0.196$, F = 77.214***
	Ratios of productivity	0.200	3.626***	$R^2 = 0.040$, F = 13.151***

Note *p < 0.05, **p < 0.01, *** < 0.001

negative (−) effect on Ratios of Profitability (t = −2.636, p = 0.009) and Ratios of Productivity (t = −2.718, p = 0.007). However, the amount of technology transaction did not have a significant effect on Ratios of Growth (t = 1.421, p = 0.156).

As a result of analysing the effect of technology readiness level on technology transfer performance, technology readiness level has a significant positive (+) effect on Ratios of Profitability (t = −3.732, p = 0.000), Ratios of Growth (t = 8.787, p = 0.000) and Ratios of Productivity (t = 3.626, p = 0.000) (Table 4).

4.2 Moderating Effect Model

As a result of analysing the moderating effect of potential absorptive capacity on the relationship between technology value and technology transfer performance, the relationship between technology development stage and ratios of profitability (t = 2.176, p = 0.030) showed a significant effect (Table 5).

As a result of analysing the moderating effect of realized absorptive capacity on the relationship between technological value and technology transfer performance, the relationship between TRL and ratios of profitability (t = 1.362, p = 0.174) was excluded, and it was discovered that all the relationships between technology value and technology transfer performance were significant (Table 6).

Table 5 Moderating effects of potential absorptive capacity

Independent variable	Moderating variable	Dependent variable	β	t	p
Technology transfer amount	Potential absorptive capacity	Ratios of profitability	0.100	1.232[a]	0.219
		Growth rates	0.145	1.772[a]	0.077
		Ratios of productivity	−0.115	−1.428[a]	0.154
Technology readiness level	Potential absorptive capacity	Ratios of profitability	0.558	1.517[b]	0.130
		Growth rates	0.732	2.176* [b]	0.030
		Ratios of productivity	0.311	0.852[b]	0.395

Note *p < 0.05, **p < 0.01, *** < 0.001
[a]Technology Transfer Amount * Potential Absorptive Capacity
[b]Technology Readiness Level * Potential Absorptive Capacity

Table 6 Moderating effects of realized absorptive capacity

Independent variable	Moderating variable	Dependent variable	β	t	p
Technology transfer amount	Realized absorptive capacity	Ratios of profitability	0.145	2.037* [a]	0.042
		Growth rates	0.236	3.219** [a]	0.001
		Ratios of productivity	0.254	3.795*** [a]	0.000
Technology readiness level	Realized absorptive capacity	Ratios of profitability	0.790	2.963** [b]	0.003
		Growth rates	0.359	1.362[b]	0.174
		Ratios of productivity	1.228	5.004*** [b]	0.000

Note *p < 0.05, **p < 0.01, *** < 0.001
[a]Technology Transfer Amount * Realized Absorptive Capacity
[b]Technology Readiness Level * Realized Absorptive Capacity

5 Conclusions

5.1 Summary and Discussion of Results

We conducted this study to determine the effect of technology value on the technology transfer performance of firms by setting the technology transaction amount and technology readiness level as the technical value, based on actual technology transac-

tion information. In addition, this study verified the moderating effect of absorption capacity between technology values and technology transfer performance.

The results of this study can be summarized as follows: First, the technology transaction amount has a negative effect on the technology transfer performance. Specifically, it has a negative (−) effect on profitability and productivity among corporate technology transfer performance indexes. These analytical results can be interpreted as the fact that, even if the technological transaction amount is high, it does not necessarily improve the technology transfer performance. Narin et al. [39] said that patent holdings of a company do not affect the management performance of a company. Morbey and Reithner [40] said that there is a negative relationship between R&D investment and profit margins. Technology with a high technology transaction amount is often determined by innovation or superiority of technology. In other words, if the technology is innovative or new technology, the value of technology is high, and because the amount of technology transaction is calculated on that basis, it takes a considerable amount of time until technology commercialization for the firms with the transferred technology.

Second, the technology readiness level has a positive effect on the technology transfer performance. A high technology readiness level (TRL) means that the technology completeness is high, which indicates that the company with transferred technology can use the technology to quickly enter the market as well as develop new products. This result is consistent with the results of previous studies demonstrating that the higher the technology readiness level, the greater the possibility of commercialization [7, 21, 22].

Third, in the relationship between technology value and technology transfer performance, absorption capacity has a partial moderating effect. Specifically, the potential absorption capacity has a moderating effect on the relationship between the R&D stage and the profitability of the firm. Potential absorption capacity is the ability to interpret and process externally acquired knowledge and skills [34], which can improve the interpretation and understanding of new external information and speed up R&D activities [41–43]. In other words, if potential absorption capacity is high, firms' ability to analyse internal technology is also high, so that external technology is easily understood and internalized, contributing to the firms' technology transfer performance. As mentioned earlier, technology with a high TRL alone has a positive effect on improving the technology transfer performance of firms. If the R&D concentration is high, it will contribute to the enhancement of the financial performance of firms. Also, feasible absorption capacity has a moderating effect on technological value and the technology transfer performance of firms. The feasible absorption capacity is the ability to combine and assimilate external knowledge and prior knowledge, creating new knowledge [34]. If the feasible absorption capacity is high, the time and effort required to commercialize innovative technology is reduced, positively affecting the management achievement of firms.

Taken together, technology with a higher TRL, rather than technology with a high technology transaction amount, contributes to improving the technology transfer performance of firms. In other words, technology with a high technology transaction value can be innovative technology, but it is meaningless to expect immediate tech-

nology transfer performance. Furthermore, technology absorption capacity plays an important role in maximizing the value of the transferred technology.

5.2 Theoretical and Managerial Implications

The implications of this study can be derived as follows: First, when measuring the value of technology, it is necessary to consider the monetary value of technology and the technology readiness level as complementary. The technology transaction amount is based on the monetary value of technology when technology is innovative, and its industrial ripple effect is substantial, technology value is given a high measure and transaction amount is also high. However, as indicated by the results of this study, it does not have a significant effect on improving the technology transfer performance of firms. On the contrary, it was found that the higher the technology readiness level, the greater the effect on the technology transfer performance. Firms are recognized in the relevant markets by their innovative technologies, and they seek to achieve continuous growth through innovative technologies. Therefore, when measuring the value of technology, it is necessary to complement and utilize the technology while considering its monetary value and technology readiness level.

Second, to improve the technology transfer performance of firms, firms should consider their technology absorption capacity. Even if technology with a high technology value is introduced, if the technology absorption capacity of the firm is low, the technology cannot fully realize its potential value. Therefore, when considering the introduction of external technology, the firm's technology absorption capacity should be considered.

Third, when calculating the technology transaction amount, it is necessary to take into account the absorption capacity of the firms with transferred technology. The technology valuation amount, which is the basis for calculating the value of technology transactions, does not take into account the competence of the firm that conducts business with that technology when calculating the monetary value of the technology. In other words, technology with a high technology transaction amount is likely to be calculated considering the potential value of technology.

This study is meaningful because it has suggested empirical strategic directions to improve the technology transfer performance of firms with transferred technology, through empirical verification based on actual technology transfer information.

5.3 Limitations and Further Research

This study empirically validated the effect of public technology value on business performance from the perspective of demand Companies, but failed to fully review the results of this study because of the lack of prior research on technology transfer performance from the perspective of demand enterprises. We expect that future

research will provide a broader understanding of the importance of technology value for public technology transfer and enhancement if the limitations of this study are supplemented.

References

1. Jamison, D.W., Jansen, C.: Technology transfer and economic growth. Ind. High. Educ. **15**(3), 189–196 (2001)
2. Min, J.W., Kim, Y.J.: A study of success factors in public technology transfer: the implications of licensee's motivation. J. Intellect. Prop. **10**(2), 225–256 (2015)
3. Sung, T.E., Kim, D.S., Jang, J.M., Park, H.W.: An empirical analysis on determinant factors of patent valuation and technology transaction prices. J. Korea Technol. Innov. Soc. **19**, 254–279 (2016)
4. Bozeman, B., Crow, M.: Technology transfer from US government and university R&D laboratories. Technovation **11**(4), 231–246 (1991)
5. Capon, N., Glazer, R.: Marketing and technology: a strategic co-alignment. J. Mark. **51**(3), 1–14 (1987)
6. Korea Institute for Advancement of Technology (KIAT): research on the status of technology transfer project. Public research institute (2015)
7. Park, J.O., Youn, S.J., Park, B.S.: Commercialization success factors of transfer technology from public R&D and enhancing performance. J. Korea Technol. Innov. Soc. **18**, 28–48 (2015)
8. Boer, F.P.: The Valuation of Technology: Business and Financial Issues in R&D. Wiley, New York (1999)
9. Thamhain, H.J.: Management of Technology, pp. 254–277. Willey, New York (2005)
10. Noori, H.: Managing the Dynamics of New Technology: Issues in Manufacturing Management. Prentice Hall, Englewood Cliffs (1990)
11. Smith, G.V., Parr, R.L.: Valuation of Intellectual Property and Intangible Assets. Wiley, New York (2000)
12. Hellmann, T.: The role of patents for bridging the science to market gap. J. Econ. Behav. Organ. **63**(4), 624–647 (2007)
13. Sadin, S.R., Povinelli, F.P., Rosen, R.: The NASA technology push towards future space mission systems. Acta Astronaut. **20**, 73–77 (1989)
14. Mankins, J.C.: Technology readiness levels. White Paper, 6 Apr 1995
15. Ahn, E.Y., Kim, S.Y., Lee, J.W.: Technology readiness levels (TRLs) indicator development for geoscience and mineral resources R&D. Econ. Environ. Geol. **48**(5), 421–429 (2015)
16. Hwang, H.W., Kim, H.R., Chang, Y.K.: TRL impact on development schedule and cost in the aerospace project. J. Korean Soc. Aeronaut. Space Sci. **40**(3), 264–272 (2012)
17. DoD, U.S.: Technology readiness assessment (TRA) guidance. Revision Posted **13** (2011)
18. Engel, D.W., Dalton, A.C., Anderson, K.K., Sivaramakrishnan, C., Lansing, C.: Development of Technology Readiness Level (TRL) Metrics and Risk Measures. Pacific Northwest National Laboratory (2012)
19. Kim, N.G., An, B.H., Lee, H.S., Choi, J.H., Park, S.H., Kim, Y.S.: Implementation of TRL and TRA tools to Korean construction and transportation R&D evaluation for improving practical use. Korean J. Constr. Eng. Manag. **13**(4), 110–119 (2012)
20. Korea Evaluation Institute of Industrial Technology (KEIT): TRL evaluation indicator for industrial strategic technology (2009)
21. Ettlie, J.E.: The commercialization of federally sponsored technological innovations. Res. Policy **11**(3), 173–192 (1982)
22. Lee, Y.D.: A study of R&D strategy-environmental factors-performances of the academia in the information and telecommunication industry: an analysis of ITRC projects. J. Korea Technol. Innov. Soc. **11**, 431–449 (2008)

23. Teece, D.J., Pisano, G., Shuen, A.: Dynamic capabilities and strategic management. Strat. Manag. J. **18**(7), 509–533 (1997)
24. Ernst, H.: Patenting strategies in the German mechanical engineering industry and their relationship to company performance. Technovation **15**(4), 225–240 (1995)
25. Bak, S.H., Lee, C.G., Seo, C.S.: The effects of green technology patent on the financial performance of specialized green enterprises. J. Korea Technol. Innov. Soc. **16**, 724–753 (2013)
26. Yam, R.C., Guan, J.C., Pun, K.F., Tang, E.P.: An audit of technological innovation capabilities in Chinese firms: some empirical findings in Beijing, China. Res. Policy **33**(8), 1123–1140 (2004)
27. Branch, B.: Research and development activity and profitability: a distributed lag analysis. J. Polit. Econ. **82**(5), 999–1011 (1974)
28. Kamiyama, S., Sheehan, J., Martinez, C.: Valuation and exploitation of intellectual property. STI Working Paper Series, OECD DSTI/DOC (2006)
29. Sohn, D.W.: A study of knowledge transfer effects in Korean venture startups: the role of knowledge origins, absorptive capacity, government, and venture capital. J. Technol. Innov. **18**, 21–51 (2010)
30. Cohen, W.M., Levinthal, D.A.: Absorptive capacity: a new perspective on learning and innovation. Adm. Sci. Q. **35**(1), 128–152 (1990)
31. Zahra, S.A., Nielsen, A.P.: Sources of capabilities, integration and technology commercialization. Strat. Manag. J. **23**(5), 377–398 (2002)
32. Escribano, A., Fosfuri, A., Tribó, J.A.: Managing external knowledge flows: the moderating role of absorptive capacity. Res. Policy **38**(1), 96–105 (2009)
33. Zhou, K.Z., Wu, F.: Technological capability, strategic flexibility, and product innovation. Strat. Manag. J. **31**(5), 547–561 (2010)
34. Zahra, S.A., George, G.: Absorptive capacity: a review, reconceptualization, and extension. Acad. Manag. Rev. **27**(2), 185–203 (2002)
35. Lenox, M., King, A.: Prospects for developing absorptive capacity through internal information provision. Strat. Manag. J. **25**(4), 331–345 (2004)
36. Schmidt, T.: Knowledge flows and R&D co-operation: firm-level evidence from Germany. ZEW-Centre for European Economic Research Discussion Paper (05-022) (2005)
37. Mosakowski, E.: Organizational boundaries and economic performance: an empirical study of entrepreneurial computer firms. Strat. Manag. J. **12**(2), 115–133 (1991)
38. Yeoh, P.L., Roth, K.: An empirical analysis of sustained advantage in the US pharmaceutical industry: impact of firm resources and capabilities. Strat. Manag. J. **20**(7), 637–653 (1999)
39. Narin, F., Noma, E., Perry, R.: Patents as indicators of corporate technological strength. Res. Policy **16**(2–4), 143–155 (1987)
40. Morbey, G.K., Reithner, R.M.: How R&D affects sales growth, productivity and profitability. Res.-Technol. Manag. **33**(3), 11–14 (1990)
41. Daft, R.L., Lengel, R.H.: Organizational information requirements, media richness and structural design. Manag. Sci. **32**(5), 554–571 (1986)
42. Egelhoff, W.G.: Information-processing theory and the multinational enterprise. J. Int. Bus. Stud. **22**(3), 341–368 (1991)
43. Rosenkopf, L., Nerkar, A.: Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. Strat. Manag. J. **22**(4), 287–306 (2001)

Author Index

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 850,
https://doi.org/10.1007/978-3-030-26428-4