Studies in Computational Intelligence 1074

Roger Lee *Editor*

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

Springer

Studies in Computational Intelligence

Volume 1074

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.

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

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

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 978-3-031-19603-4 ISBN 978-3-031-19604-1 (eBook)
https://doi.org/10.1007/978-3-031-19604-1

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

Preface

The purpose of the 23rd ACIS International Summer Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD2022-Summer) held during July 4–6, 2022, in Kyoto City, 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 Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing discussed a wide range of issues with significant implications, from Artificial Intelligence to Communication Systems and Networks, Embedded Systems, Data Mining and Big Data, Data driven business models, Data privacy and security issues, etc.

This publication captures 15 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 the chapter "A-IDE: A Non-intrusive Cross-Platform Development Environment for AVR Programming in Assembly", Antoine Bossard proposes a A-IDE, a non-intrusive cross-platform development environment for AVR programming in assembly, importantly retaining the pedagogical principles of popular development environments, such as Arduino's, support only high-level languages like C and C++. They show that the proposed system outperforms existing ones from the disk space requirement point of view and that it is on par with those with respect to memory (RAM) utilization.

In the chapter "Develop a Horizontal Virtual Frame by Adding Field of View Restrictions to Reduce VR Sickness", Hexi Huang presents a method for preventing VR sickness based on the previous development of the virtual horizontal frame. On the basis of confirming that the virtual horizontal frame has a good effect on VR sickness, he further studied the effect of restricting part of the field of view.

In the chapter "Represent Score as the Measurement of User Influence on Twitter", Yuto Noji, Ryotaro Okada and Takafumi Nakanishi present a novel method for quantifying a user's ability to spread information based on the number of Retweets (RTs) and Likes they receive on Twitter. In this novel method, they propose a method for extracting indicators that show the diffusion power, not of tweets alone, but users as

a unit, by measuring the ratio of the number of RTs and Likes based on the number of RTs and Likes of users in the past.

In the chapter "The Experience of Developing a FAT File System Module in the Rust Programming Language", Shuichi Oikawa describes the experience of developing the FAT file system as a kernel module in Rust employing Rust for Linux as a basis of the development. They performed the experiments to measure its execution costs, and found that its performance is comparable with the original FAT file system written in C.

In the chapter "Factors Influencing Consumers' Online Grocery Shopping Under the New Normal", Satoshi Nakano aims to assess the psychological factors that influence consumers' actual online grocery purchases under the new normal by combining purchase panel data and survey data. This study confirms that online grocery purchase amount is affected by traditional utilitarian channel choice factors including perceived risk, search cost, price-consciousness and quality-consciousness.

In the chapter "Skeletal Muscle Segmentation at the Third Lumbar Vertebral Level in Radiotherapy CT Images", Xuzhi Zhao, Haizhen Yue, Yi Du, Shuang Hou, Weiwei Du and Yahui Peng propose a computer algorithm to segment skeletal muscles at the third lumbar vertebral (L3) level in radiotherapy computed tomography (CT) images. Included in the study are 20 patients who were diagnosed with rectal cancer and their pelvic CT images were acquired with a radiotherapy CT scanner.

In the chapter "Speed-up Single Shot Detector on GPU with CUDA", Chenyu Wang, Toshio Endo, Takahiro Hirofuchi and Tsutomu Ikegami implemented a Single Shot Multibox Detector (SSD) using GPU with CUDA. They have improved the object detection speed of SSD, which is one of the most regularly used object detection frameworks.

In the chapter "An Empirical Study on the Economic Factors Affecting on the Export of Defense Industry Using Hofstede's Culture Dimension Theory", Taeyeon Kim, Dongcheol Kim, Jaehwan Kwon and Gwangyong Gim propose the necessity of considering the cultural factors of the countries listed to export in order to establish a defense export marketing strategy effectively. To support the proposal, the defense export marketing strategies and purchasing models of the other countries, that are not clearly established at present, were proposed and partially verified in this thesis by conducting statistical analysis using data related to defense, national and military size of each country in the world.

In the chapter "A Study on the Application of Blockchain Technology in Non-governmental Organizations", Minwoo Lee, Saeyeon Lee, Heewon Lee and Gwangyong Gim use block chain technology to study the market trend of Non-Profit Organizations (NGOs). They present the implications of data for introducing and utilizing blockchain technology to non-profit organizations (NGOs) in the future and discussed the limitations of this study and future research tasks.

In the chapter "Confidential Documents Sharing Model Based on Blockchain Environment", Sung-Hwa Han proposes a model that can share confidential documents in a blockchain environment. The proposed confidential document sharing model operates on the blockchain platform and has the advantage of not using a key management server.

In the chapter "Distance Based Clustering in Wireless Sensor Network", Joong Ho Lee proposes a single unit distance based two-hop clustering method for clustering and analyzes the energy persistence of sensor nodes in the formed clusters. The proposed cluster formation method can efficiently manage operational energy by improving the non-uniformity of cluster groups. And their paper shows comparison results between the conventional and proposed schemes of clustering algorithms.

In the chapter "Study on OSINT-Based Security Control Monitoring Utilization Plan", DAIN Lee and Hoo-Ki Lee present the definition and concept of OSINT, and utilization plans in security monitor including characteristics are studied and presented, and OSINT information collection tools are identified and analyzed. Through this, OSINT-based security monitor utilization plans and considerations are presented.

In the chapter "A Study on the Strategy of SWOT Extraction in the Metavers Platform Review Data: Using NLP Techniques", Jina Lee, Euntack Im, Inmo Yeo and Gwangyong Gim propose a framework that addresses the major shortcomings of traditional SWOT analysis through NLP based on review data from Metaverse applications. The framework presented in this study presents ways to establish business strategic tools using secondary data and provides implications for important factors in building a Metaverse application environment.

In the chapter "Does Facial Expression Accurately Reveal True Emotion? Evidence from EEG Signal", Huy Tung Phuong, Yangyoung Kun, Jisook Kim and Gwangyong Gim conduct an experiment to collect facial expression data and EEG signals of the participant. The results indicate that not all kinds of basic emotions were expressed in facial expressions in the experiment environment. The study shows the problems faced by facial emotion recognition systems and proposes future works to improve the efficiency of those systems.

In the chapter "NSGA-II-AMO: A Faster Genetic Algorithm for QWSCP", Zehui Feng, Bei Wang, Mingjian Chen and Qi Chen propose an improved NSGA-II-AMO algorithm, which uses an adaptive mutation operator to complete the mutation process when generating children in the NSGA-II algorithm, so that the mutation probability adaptively changes with different distributions of data, effectively improving the search speed at the beginning of population iteration.

They compare the NSGA-II-AMO algorithm cross-sectionally with the PSO algorithm, the NSGA-II algorithm and the ABC algorithm, and the adaptive mutation operator achieves improved convergence with guaranteed distributivity.

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.

Mount Pleasant, USA Roger Lee
June 2022

Contents

Contributors

Antoine Bossard Graduate School of Science, Kanagawa University, Hiratsuka, Japan

Mingjian Chen College of Computer Science and Technology, Zhejiang University, Hangzhou, China

Qi Chen College of Computer Science and Technology, Zhejiang University, Hangzhou, China

Weiwei Du Department of Information Science, Kyoto Institute of Technology, Kyoto, Japan

Yi Du Department of Radiation Oncology, Peking University Cancer Hospital and Institute, Beijng, China

Toshio Endo Tokyo Institute of Technology, Tokyo, Japan

Zehui Feng Polytechnic Institude, Zhejiang University, Hangzhou, China

Gwangyong Gim Department of Business Administration, Soongsil University, Seoul, South Korea

Sung-Hwa Han Depatment of Information Security, Tongmyong University Busan, Busan, South Korea

Takahiro Hirofuchi National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Shuang Hou School of Electronic and information Engineering, Beijing Jiaotong University, Beijng, China

Hexi Huang Department of Information Science, Aichi Institute of Technology, Toyota, Japan

Tsutomu Ikegami National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Euntack Im Department of Business Administration, Soongsil University, Seoul, South Korea

Dongcheol Kim Department of IT Policy and Management, Soongsil Univ. Seoul, Seoul, South Korea

Jisook Kim Management Planning HQ, Soongsil University, Seoul, South Korea

Taeyeon Kim Department of Business Administration, Soongsil University Seoul, Seoul, South Korea

Yangyoung Kun Department of IT Policy and Management, Soongsil University, Seoul, South Korea

Jaehwan Kwon Department of IT Policy and Management, Soongsil Univ. Seoul, Seoul, South Korea

Dain Lee Department of T&D Security, Daejeon-Si, South Korea

Heewon Lee Department of IT Policy and Management, Soongsil University Seoul, Seoul, South Korea

Hoo-Ki Lee Department of Cyber Security Engineering, Konyang University Nonsan-Si, Nonsan-Si, South Korea

Jina Lee Department of Business Administration, Soongsil University, Seoul, South Korea

Joong Ho Lee Depatment of AI, Yongin University, Yongin-Si, South Korea

Minwoo Lee Department of IT Policy and Management, Soongsil University Seoul, Seoul, South Korea

Saeyeon Lee Department of IT Policy and Management, Soongsil University Seoul, Seoul, South Korea

Takafumi Nakanishi The Tokyo Foundation for Policy Research, Department of Data Science, Musashino University, Tokyo, Japan

Satoshi Nakano Faculty of Economics, Meiji Gakuin University, Tokyo, Japan

Yuto Noji The Tokyo Foundation for Policy Research, Department of Data Science, Musashino University, Tokyo, Japan

Shuichi Oikawa School of Industrial Technology, Advanced Institute of Industrial Technology, Tokyo, Japan

Ryotaro Okada The Tokyo Foundation for Policy Research, Department of Data Science, Musashino University, Tokyo, Japan

Yahui Peng School of Electronic and information Engineering, Beijing Jiaotong University, Beijng, China

Huy Tung Phuong Department of Business Administration, Soongsil University, Seoul, South Korea

Bei Wang College of Computer Science and Technology, Zhejiang University, Hangzhou, China

Chenyu Wang Tokyo Institute of Technology & National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Inmo Yeo Samsung Town Finance Center, Samsung Securities Co., Ltd., Seoul, South Korea

Haizhen Yue Department of Radiation Oncology, Peking University Cancer Hospital and Institute, Beijng, China

Xuzhi Zhao School of Electronic and information Engineering, Beijing Jiaotong University, Beijng, China

A-IDE: A Non-intrusive Cross-Platform Development Environment for AVR Programming in Assembly

Antoine Bossard

Abstract Minimalistic devices such as microcontrollers make a large part of the Internet of things (IoT), for instance to build sensors and sensor networks. Due to severely limited resources, especially memory, microcontroller programming faces various restrictions, such as hampered graphical data processing. In order to circumvent those weaknesses, direct programming of these devices with the assembly language has proven effective. Because popular development environments, such as Arduino's, support only high-level languages like C and C++, we propose in this paper A-IDE, a non-intrusive cross-platform development environment for AVR programming in assembly, importantly retaining the pedagogical principles of those existing popular solutions. Nowadays, microcontroller programming has indeed a high educative value: refer, for example, to the BBC micro:bit and Arduino Uno, to only cite a few. It is then quantitatively shown that the proposed system outperforms existing ones from the disk space requirement point of view and that it is on a par with those with respect to memory (RAM) utilisation.

Keywords Microcontroller · Sensor · Atmel · Microchip · Arduino · 8-bit · ASM · IoT

1 Introduction

Sensors and sensor networks are an important part of the Internet of things (IoT). Sensors are often minimalistic devices based on cost-effective chips such as, typically, microcontrollers. In order to reduce their cost as well as their power consumption, such particular hardware comes with severely limited resources, especially memory. Hence, microcontroller programming has to do with various restrictions, such as the absence of libraries, let alone of an operating system. In order to circumvent those inherent weaknesses, direct programming of these devices with the assembly language has been shown to be effective [1, 3].

A. Bossard (✉)
Graduate School of Science, Kanagawa University, 2946 Tsuchiya, Hiratsuka 259-1293, Japan
e-mail: abossard@kanagawa-u.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_1

Microcontrollers based on the AVR architecture have proven popular over the years, notably due to the widespread usage of the Arduino Uno boards equipped with the ATmega328P microcontroller of Microchip Technology [2, 11]. The AVR instruction set architecture (ISA) relies on the reduced instruction set computer (RISC) principle; it is an 8-bit architecture [12]. (It can be noted though that a 32-bit version, AVR32, which is only partly related to AVR, did also exist but is not supported any more.)

In addition, popular development environments for microcontrollers support high-level languages such as C/C++ (e.g. the Arduino development environment [8]), Python (e.g. Python Editor for BBC micro:bit [5]) and even visual programming (e.g. Scratch for BBC micro:bit [9]). Combined with an accessible user interface, these make performant solutions for general programming educational purposes. However, they are less fit for, say, computer architecture lectures of undergraduate computer science curricula. In this case, a precise understanding of hardware units and their operation is required, which is why directly relying on the assembly language is definitely meaningful in such a context [4, 6, 15].

The objective of this research project is thus to provide a non-intrusive, cross-platform development environment for AVR programming in assembly which features a high accessibility so as to be suitable for educational environments and applications. The proposed development environment is neither just a front-end nor just a code editor, it is a combination of both, and with additional, specific features for AVR programming in assembly. It is an integrated development environment (IDE). We have called it A-IDE.

The rest of this paper is organised as follows: preliminary information is given in Sect. 2. The proposal is described in detail in Sect. 3 and quantitatively evaluated in Sect. 4. Finally, Sect. 5 concludes this paper.

2 Preliminaries

AVR programming relies in general on the tool chain provided by elements of the GNU Binutils collection, namely the assembler `avr-as` from the GNU `as` assembler collection, the linker `avr-ld` from GNU `ld`, the machine code translator `avr-objcopy` from GNU `objcopy` and the machine code uploader `avrdude`. In addition, A-IDE relies on the machine code size calculation utility `avr-size` from GNU `size`.

For the sake of simplicity, portability and even robustness, these tools of the GNU Binutils collection are purposefully not bundled with A-IDE, although they could easily be, and so it is required to have them installed beforehand on the system. In addition, the operating system driver to communicate with the microcontroller (USB serial driver) also needs to be separately installed. These two prerequisites can be easily cleared by installing, for example, the Arduino IDE, or they can be addressed directly by manually installing these components in case memory usage is a concern (see Sect. 4).

3 System Description

3.1 The User Interface

Because education is one important objective of microcontroller programming, A-IDE features a user-friendly interface suited for this pedagogical matter as shown by its proximity with, for instance, Arduino's integrated development environment and Python Editor for micro:bit, and by the educational projects records of these two sample development solutions. As illustrated in Fig. 1, A-IDE's interface consists of a menu bar, a tool bar, an editor panel and an output panel.

The menu bar includes conventional features to load (e.g. open, save, reload) and edit (e.g. undo, copy, paste, find) source code files.

The tool bar shows the current line number, that is that of the caret, the name of the file being edited and an indicator of the saved/unsaved state thereof, space for an information message to show, for instance, the size of the generated machine code, and the build and upload buttons (see Sect. 3.3).

The editor panel is where the edition of the source code file happens. This component is at the core of the proposal and as such it is described in detail below in Sect. 3.2.

Finally, the output panel, which can be hidden if so wished by the user, displays build- and output-related information.

So, a simple, conventional and intuitive interface. Note that widgets (controls) native to the operating system that runs A-IDE are used to build the interface so as to flatten the learning curve of the user. The sizes of the panels, and thus of the main window, can be freely adjusted.

Figure 1 illustrates the case of the Microsoft Windows operating system. A similar illustration this time in the case of the Linux operating system with the Ubuntu distribution running on the Windows Subsystem for Linux (WSL) is given in Fig. 5 in appendix.

3.2 The Edition Component

Syntax colouring has been realised with the `color:text%` class of Racket's `framework` library and with the lexer feature provided by the `parser-tools` library. Assembler directives, instructions, registers, numeric values, labels and comments (both single and multiline) are automatically detected and emphasised (refer to Fig. 1).

The conventional text edition operations have been implemented as well: for example, text selection for the cut, copy and paste trio, the undo and redo duo, and a bidirectional text search feature. In addition, deletion of the current line and deletion of the rest of the current line are two other editing features provided in the editor panel. The number of the line where the caret is located is shown in the tool bar

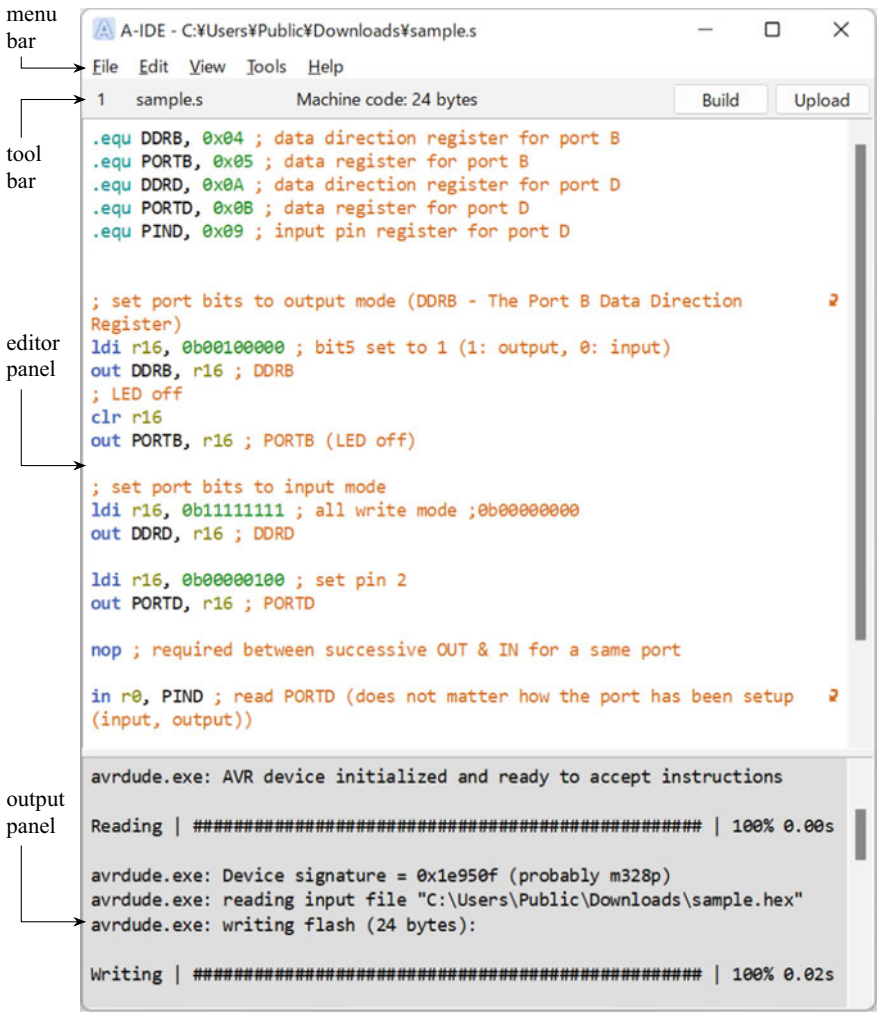


Fig. 1 An overview of the accessible user interface of A-IDE. This screenshot illustrates the case of the Microsoft Windows operating system

as presented earlier. This is an important feature especially when debugging: error messages from the assembler usually mention the number of the line that includes faulty code.

The editor component of A-IDE also features optional automatic line wrapping (i.e. soft wrap, which is exemplified in Fig. 1 by the small arrows on the right hand side of the editor panel). And, it has extensive Unicode support.

3.3 The Build and Upload Component

3.3.1 Preferences

First and foremost, the tool chain that was described in Sect. 2 can be configured in the preferences dialog of A-IDE; refer to Fig. 2. The path of each of all the components of the tool chain (i.e. `avr-as`, `avr-ld`, `avr-objcopy`, `avr-size` and `avrdude`) is specified, and either highlighted in green or red to indicate a successful setup or not, respectively. The user is given the possibility to use a file selection dialog to avoid typing in the full path. It is recalled that accessibility is essential to A-IDE.

In addition, the user can select which microcontroller is targeted, for instance the ATmega328P chip, and which programmer to use. Once again for accessibility reasons, a list of supported devices is provided, in other words, the user is not required to type in the corresponding string constant (value). Of course, the connection port (e.g. COM port on Microsoft Windows) of the device can also be declared, albeit manually this time.

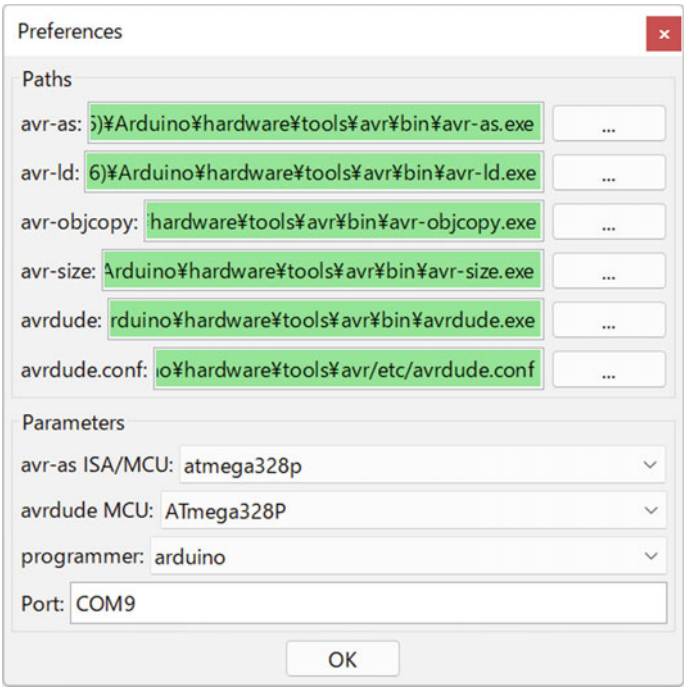


Fig. 2 The preferences dialog of A-IDE, which is used to configure, when necessary, the build and upload tool chain

3.3.2 Additional Technical Details

The build and upload process is conducted asynchronously: a second thread is started to refrain from freezing the main thread, which is notably responsible for the responsiveness of the user interface. This is thus critical so as to retain system usability. Technically, this is achieved with the `thread` function of Racket.

In addition, the build and upload commands, run in this second thread, are also called in a non-blocking manner in order to retrieve their respective output data as they are emitted. Otherwise, such data would be available only once the command has finished its execution. As a result, it is possible to track in detail the execution progress of the build and upload commands. Technically, this is achieved with the `subprocess` function of Racket, and its siblings such as `subprocess-wait` and `subprocess-status`.

Besides, the standard output and standard error streams are merged with the `merge-input` function and are thus both redirected to A-IDE, the parent (calling) process. Their content is displayed in the output panel.

Furthermore, the upload process can be aborted since `avrdude` could enter a "retry loop" upon failure, for instance when the connection to the microcontroller cannot be established. The flow of the build and upload thread is illustrated in Fig. 3; each element of the tool chain is run asynchronously with this thread, which is thus operating in a non-blocking mode.

Regarding the building process, any `.eeprom` section as output by the linker `avr-ld` is excluded with the translation utility `avr-objcopy` as such a section is not bound for the program memory in flash, obviously. But since programming directly in assembly, the presence of an `.eeprom` section is unlikely, unless so desired by the user.

3.4 System Requirements

As mentioned previously, the realised system is cross-platform: thanks to the Racket language and its portable libraries, A-IDE can be seamlessly used on Microsoft Windows, Apple MacOS and other Unix-based operating systems. Refer, for example, to Figs. 1 and 5 for an illustration. In addition, both 32-bit and 64-bit systems are supported. There are no specific CPU, RAM and so on requirements: a computer that is modern enough will most likely do. For instance, only about 64.2 MB of free disk space are required in the case of Microsoft Windows to run A-IDE. Additional details, regarding system requirements, which we obtained after evaluation of the proposal are given next in Sect. 4.

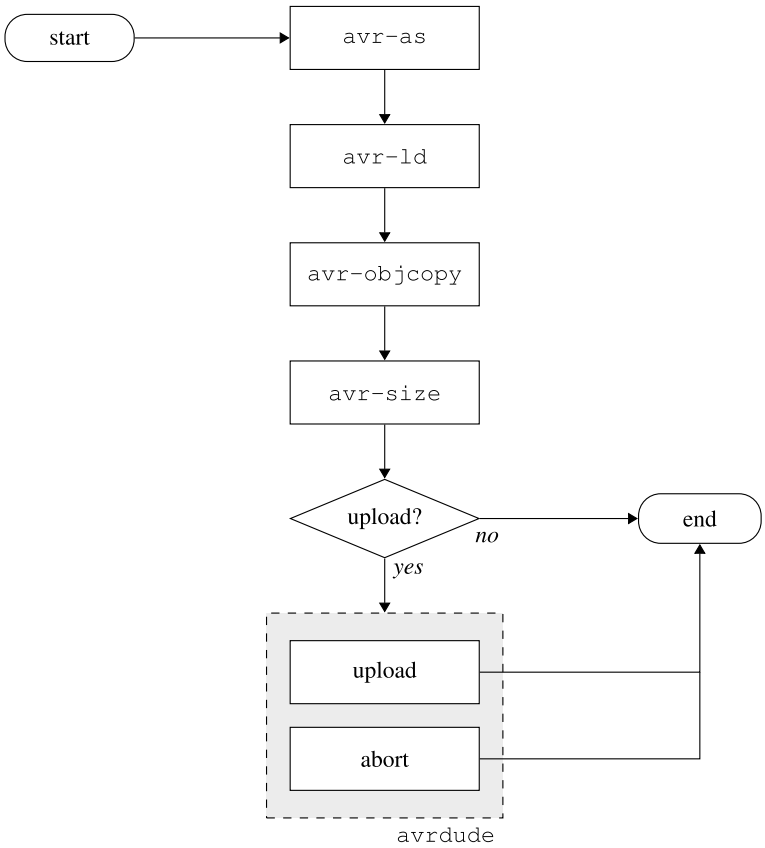


Fig. 3 Flow of the build and upload thread. Each element of the tool chain is run asynchronously with this thread, which is thus non-blocking

4 Comparison to Related Systems and Evaluation

We use native controls unlike the Arduino IDE and Python Editor for micro:bit as such controls are desirable to flatten the learning curve of the user: the system user is already well accustomed to them. More generally, native controls are expected in various scenarios [7, 10].

The installation package of A-IDE (i.e. the corresponding ZIP file) for Microsoft Windows takes only 36.0 MB of disk space, and only 64.2 MB of disk space once installed (i.e. unzipped) on this operating system. This is to be compared with the 200 MB of the installation package of the Arduino IDE (in the ZIP format) and

the 539 MB of disk space taken once installed (for version 1.8.19). This is also to be compared with the 933 MB required let alone by the installation package of Microchip Studio (for the offline installer of version 7.0.2542) and the very large amount of required disk space: 6 GB [13]. It should be noted as well that the size of the installation package also directly impacts the data transfer costs over the network as it is very likely that such data would be retrieved directly from the Internet.

Once installed, we have calculated that Microchip Studio (formerly Atmel Studio) and bundled software take 1.77 GB for XC8, 750 MB for Microchip Studio (including the tool chain), 234 MB for Visual Studio components, 13.5 MB for another Microchip folder and 58.8 MB for MSBuild. So, about 2.83 GB in total. It is only an approximation since it is difficult to track all the components installed by Microchip Studio on the whole operating system. Note that for a fair comparison, Microchip Studio was installed only for the AVR architecture, that is, not including the UC3 and SAM architectures, and without any extension.

Yet, it should be noted that, as explained in Sect. 2, A-IDE does not include the tool chain, unlike the Arduino IDE and Microchip Studio. So, we add the size of the tool chain to A-IDE for a fair comparison: the tool chain takes 54 MB of the installation package of the Arduino IDE and 217 MB once installed. Besides, in the case of Microchip Studio, it consumes 148 MB once the software has been installed. So, we can consider that the size of the installation package of A-IDE plus the tool chain is at most $36 + 54 = 90$ MB, and that once installed, A-IDE together with the tool chain take at most $64.2 + 217 = 281.2$ MB. The comparison of these actual disk space requirements is illustrated in Fig. 4 and summarised in Table 1. Note that the USB serial communication driver files for the operating system, which are neither included in A-IDE, are of negligible size and thus not taken into account here.

Of course, Microchip Studio is a development environment packed with numerous features. However, A-IDE is cross-platform whereas Microchip Studio supports only the Microsoft Windows operating system [13]. And, although Microchip Studio does enable programming directly with the assembly language, its documentation rather clearly shows that it favours programming in C/C++ instead [14]. As a result, A-IDE is more suitable for the educational purposes mentioned previously in introduction.

Table 1 Comparison of the actual disk space required for the installation package and after installation of A-IDE (with the tool chain), Arduino IDE and Microchip Studio

	A-IDE + tool chain (in megabytes)	Arduino IDE (in megabytes)	Microchip Studio (in megabytes)
Installation package	90	200	933
After installation	281	548	2826
Total	371	748	3759

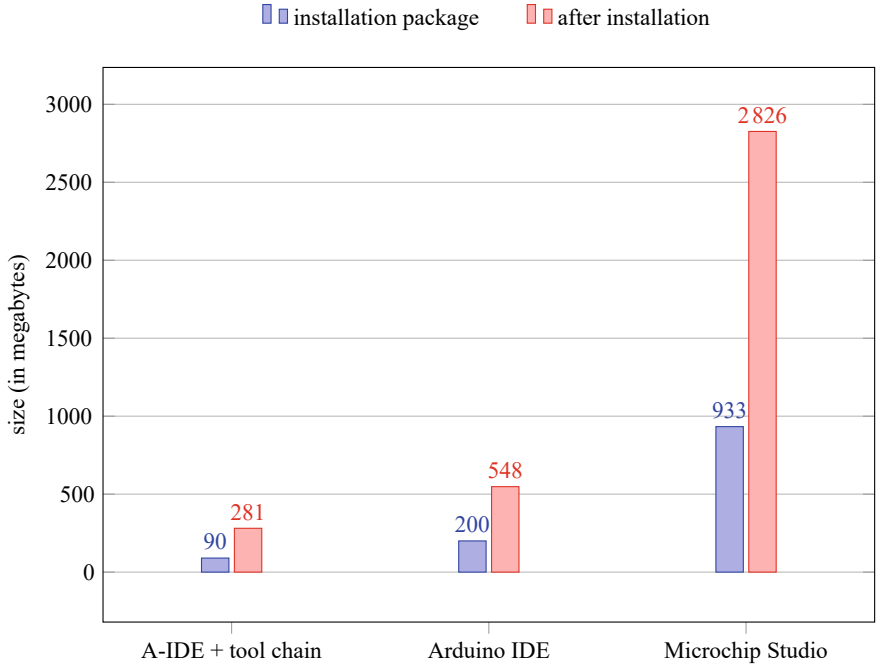


Fig. 4 Disk space required for the installation package and after installation of A-IDE (with the tool chain), Arduino IDE and Microchip Studio (in megabytes)

After file edition and running a build and upload process, the RAM memory utilisation of A-IDE stays at 333.8 MB on Microsoft Windows. This is to be compared with the 221.2 MB taken in the RAM by the Arduino IDE after similar operations on the same operating system. Hence, the RAM utilisation of our proposal is slightly more demanding: this is due to the Racket libraries that are required to run the executable file. The Java machine code that is required by the Arduino IDE consumes less RAM in this case. But considering the amount of available RAM on a typical modern computer—in the order of several gigabytes—, this difference of approximately 113 MB remains negligible.

Although an advanced text editor such as Emacs could provide syntax colouring and the integration of system commands and their input and output streams, this would require a more or less difficult setup and, besides, such a text editor can hardly provide all the specialised features of A-IDE, like the facilitated build and upload process configuration and the build information message which notably shows the size of the produced machine code after build.

5 Conclusions

The Internet landscape has evolved significantly with the advent of the Internet of things, with sensors being one category of such devices. They are in general minimalistic devices to reduce costs as much as possible. As a result, their programming faces several restrictions, severe memory limitations being one notable example. Direct programming of the chip with the assembly language is thus important to keep memory utilisation at a minimum.

In this research, we have introduced A-IDE, a non-intrusive and cross-platform development environment for programming the AVR architecture in assembly. Related systems often favour high-level languages instead, such as C/C++ for the Arduino IDE. We have quantitatively shown the comparatively low requirements of A-IDE, and we have qualitatively demonstrated its practicability, accessibility and cross-platform ability. These results have shown that A-IDE enables a scenario, programming within an IDE directly in assembly, which would be otherwise significantly more costly—Microchip Studio: more intrusive, resource expensive, platform specific (no portability), closed-source software.

Regarding future works, adding a serial monitor feature to further improve interfacing with the microcontroller is a meaningful objective. In addition, support for common debugging features of microcontrollers such as JTAG and debugWIRE is yet another possible research direction. Finally, integration of user-friendly library dependency support could also be investigated.

Acknowledgements The author is sincerely grateful towards the reviewers for their comments and suggestions. This research was partly supported by a Grant-in-Aid for Scientific Research (C) of the Japan Society for the Promotion of Science under grant no. 19K11887.

Appendix

An illustration of the user interface in the case of the Linux operating system with the Ubuntu distribution running on the Windows Subsystem for Linux (WSL) is given in Fig. 5.

Finally, the source code of A-IDE as well as binaries can be downloaded from the official homepage at https://www.sci.kanagawa-u.ac.jp/info/abossard/a-ide/.

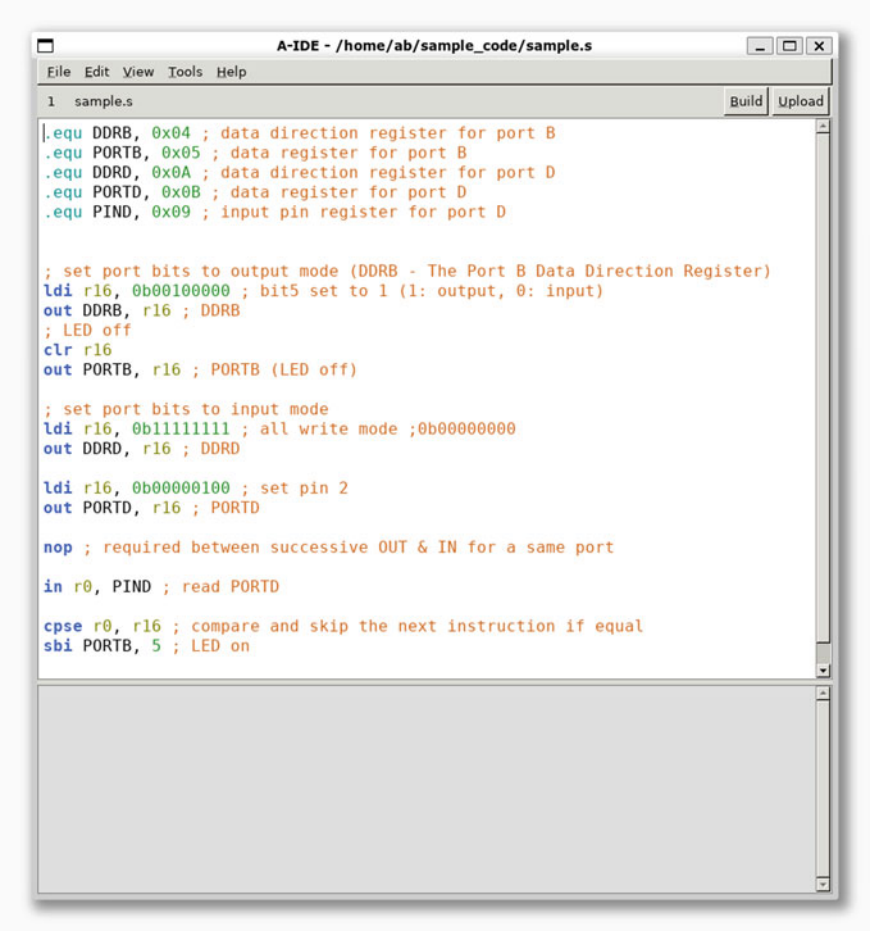


Fig. 5 An overview of the accessible user interface of A-IDE. This screenshot illustrates the case of the Linux operating system with the Ubuntu distribution running on WSL

References

1. Bossard, A.: Autonomous on-chip debugging for sensors based on AVR microcontrollers. J. Sens. Technol. **11**(2), 19–38 (2021). https://doi.org/10.4236/jst.2021.112002
2. Bossard, A.: Understanding microcontrollers -a gentle introduction to an AVR architecture-. Ohmsha, Tokyo, Japan (2021)
3. Bossard, A.: Memory optimisation on AVR microcontrollers for IoT devices' minimalistic displays. Chips **1**(1), 2–13 (2022). https://doi.org/10.3390/chips1010002
4. Chau, C.F., Fung, Y.F.: A tool for self-learning assembly language programming and computer architecture: Design and evaluation. Comput. Appl. Eng. Educ. **19**(2), 286–293 (2011). https://doi.org/10.1002/cae.20310
5. Halfacree, G.: The official BBC micro:bit user guide. Hoboken, NJ, USA (2017)

6. Hara, S., Imai, Y.: Register-transfer-level CPU simulator for computer architecture education and its quantitative evaluation. IEEJ Trans. Electron. Inf. Syst. **138**(9), 1123–1130 (2018). https://doi.org/10.1541/ieejeiss.138.1123
7. Leijen, D.: WxHaskell: a portable and concise GUI library for Haskell. In: Proceedings of the 2004 ACM SIGPLAN Workshop on Haskell, Snowbird, Utah, USA, 22 September 2004, pp. 57–68 (2004). https://doi.org/10.1145/1017472.1017483
8. Margolis, M., Jepson, B., Weldin, N.R.: Arduino Cookbook: Recipes to Begin, Expand, and Enhance Your Projects, 3rd edn. O'Reilly Media, Sebastopol, CA, USA (2020)
9. McManus, S.: Scratch Programming in Easy Steps, 2nd edn. Leamington Spa, UK (2019)
10. Meidinger, M., Ebbers, H., Reimann, C.: AeroFX – native themes for JavaFX. In: Dregvaite G., Damasevicius R. (eds.) Information and Software Technologies (Proceedings of the 21st International Conference on Information and Software Technologies; Druskininkai, Lithuania, 15–16 October 2015). Communications in Computer and Information Science, vol. 538, pp. 526–536 (2015). https://doi.org/10.1007/978-3-319-24770-0_45
11. Microchip Technology: ATmega48A/PA/88A/PA/168A/PA/328/P megaAVR data sheet (2018). DS40002061A, ISBN: 978-1-5224-3502-0
12. Microchip Technology: AVR instruction set manual (2020). DS40002198A, ISBN: 978-1-5224-5882-1
13. Microchip Technology: Microchip studio release note (2020). DS50002917C, ISBN: 978-1-5224-7063-2
14. Microchip Technology: Microchip studio user guide (2022). DS50002718E, ISBN: 978-1-5224-9776-9
15. Sondag, T., Pokorny, K.L., Rajan, H.: Frances: a tool for understanding computer architecture and assembly language. ACM Trans. Comput. Educ. **12**(4), 14:1–14:31 (2012). https://doi.org/10.1145/2382564.2382566

Develop a Horizontal Virtual Frame by Adding Field of View Restrictions to Reduce VR Sickness

Hexi Huang

Abstract VR sickness is a symptom that similar to motion sickness such as strong discomfort, nausea, dizziness, and headache, and 3D sickness when using VR. As VR becomes more and more widely used, the accompanying VR sickness also needs to be paid attention to. This research is based on the previous development of the virtual horizontal frame. On the basis of confirming that the virtual horizontal frame has a good effect on VR sickness, we will further study the effect of restricting part of the field of view. The effect was visualized by using SSQ (Simulator Sickness Questionnaire) to see the degree of sickness. Through experiments, it was found that different visual field limitations can alleviate various symptoms to different degrees. We can also formulate different visual specialization strategies according to different results of each SSQ questionnaire.

Keywords VR sickness · Virtual horizontal frame · Field of view · SSQ

1 Introduction

In recent years, virtual reality has been widely used in real estate previews, rehabilitation of sequelae caused by cerebral infarction, work education, entertainment, and advertising industries. However, as it became more widespread in the world, the drawbacks of VR have become apparent. When using VR, symptoms similar to motion sickness such as strong discomfort, nausea, dizziness, and headache, and 3D sickness may occur. These are called "VR sickness". The degree and type of sickness symptoms vary greatly from person to person, and some people do not develop the symptoms, some develop them immediately, and some people have the symptoms for a long time.

In the previous research, we approached from the visual aspect among the elements of VR sickness, and aimed to reduce sickness only by images without the need for peripheral devices or special tools other than VR. Then, in order to suppress the

H. Huang (✉)
Department of Information Science, Aichi Institute of Technology, Toyota, Japan
e-mail: b18803bb@aitech.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_2

occurrence of vection and confusion of the spatial recognition function, which are one of the causes of VR sickness, a virtual horizontal frame that "covers the peripheral visual field" and "visualizes the actual horizontal position and the inclination of one's own field of view" was developed. Post-experimental SSQ results showed a decrease in SSQ scores for all subjects wearing virtual horizontal frames. In particular, it was found that the subjects with high SSQ scores in the "no virtual horizontal frame" state were effective in reducing the feeling of discomfort and wobbling by wearing the virtual horizontal frame.

However, in the previous study, the content that verified the effect of the virtual horizontal frame was limited to the roller coaster. Therefore, we have not verified the effect on VR content that can be moved arbitrarily in 3D space by operating the controller while it is stationary, or that is performed while moving the body. We also verified the virtual horizontal frame using the horizon and the ring, but did not compare and verify it with the virtual frames of other designs. Therefore, it was not possible to determine whether the effect of suppressing VR sickness was due to the horizon placed in the center or the effect brought about by covering the peripheral visual field.

Therefore, in this research, in order to verify the effect of the virtual horizontal frame, in addition to the content that the player moves by himself/herself using the controller, the effect on the content that cannot exist in reality such as shooting games is measured. We will also develop a frame that only covers the peripheral vision and verify whether the horizon has the effect of suppressing VR sickness. We measured each of the frame that obscures the peripheral visual field, the virtual horizontal frame of the previous study, and the virtual horizontal frame of this study, and verified whether the horizon is effective.

2 Purpose of Research

2.1 Reduce VR Sickness Caused by VR Content

First of all, it is necessary to understand VR sickness in order to reduce VR sickness. The main cause of VR sickness is an external stimulus to the body or a gap between movement and visual information. VR sickness is considered as a type of agitation. In this study, VR sickness was regarded as a type of motion sickness, and the study was conducted with reference to existing preventive methods and solutions for motion sickness.

2.2 Realized by Video Processing Without Using Auxiliary Equipment

As a method of reducing VR sickness, a method of using an external stimulus to the body and a method of giving supplementary information to visual information can be considered. In this study, we reduce VR sickness by supplementing visual information with filters and video effects. As a result, it can be realized only in the environment necessary for using VR, and it can be easily introduced without the need to prepare special equipment.

2.3 Compare Virtual Frames and Verify the Effect of Horizontal Lines

In the previous study, we created a virtual horizontal frame, but we could not determine whether the effect of suppressing VR sickness was due to the horizontal line placed in the center or the effect brought about by covering the peripheral visual field. Therefore, in this study, we will develop a frame that only covers the peripheral visual field and verify whether the horizon has the effect of suppressing VR sickness. We measured each of the frame that obscures the peripheral visual field, the virtual horizontal frame of the previous study, and the virtual horizontal frame of this study, and verified whether the horizon is effective.

2.4 Prove that It Works for Various Contents

Actually, use the created virtual frame and compare the degree of sickness with the unused state. At that time, a questionnaire for quantifying sickness by a simulator called SSQ (Simulator Sickness Questionnaire) is used as an index. SSQ is performed on the subject after viewing the content in a relaxed state and without using the virtual frame, and after viewing the content with the virtual frame, and the effect of the virtual frame is proved by comparing the results. In addition, the cerebral blood flow meter is used under the same conditions, and the effect of the frame is confirmed by using not only the subjective evaluation but also the physiological index. Also, in the previous research, we experimented with one VR content, but in this research, we use various VR contents and confirm that it is effective for a wide variety of contents.

3 VR Sickness

The cause of VR sickness is thought to be the same as that of motion sickness because the symptoms of VR sickness and agitation are similar [1, 2].

One of the causes of motion sickness is the theory of sensory contradiction. There is a contradiction between the body's sense of balance information obtained by the human semicircular canals and the visual information obtained by the eyes. Motion sickness develops when they accumulate in the brain [3–5].

Since the cause of developing VR sickness is related to the cause of motion sickness, it is necessary to apply the existing mechanism for reducing the symptoms of motion sickness to VR and explore the effect in order to reduce VR sickness. VR sickness is attributed to the discrepancy between sensory and visual information from the vestibule and somatosensory, and one of the causes of VR sickness is visually induced self-motion sensation (vection) [1].

VR sickness can be suppressed by suppressing the vection that occurs when using VR content. In order to suppress vection, it is necessary to eliminate the discrepancy between visual information and spatial information obtained from the vestibule and somatosensory. There are two main methods for eliminating vection.

The first is a method that does not cause vection by making the VR environment and the natural environment the same. By transmitting visual information from the VR environment as motion information to the vestibule and somatosensory using a machine, the spatial information from the VR environment and the natural environment are matched. In addition, the spatial information can be matched by completely reproducing the natural environment on the VR environment. Reported that VR sickness was significantly reduced by creating a motorcycle in a VR environment and synchronizing the magnitude of engine sound and vibration with the image [6].

However, there are various problems in introducing a machine that reproduces the motion of VR in the natural environment. In order to use the machine, VR can be used only in a place where a sufficient range of movement is secured. Furthermore, considering the increase in the cost of introducing a machine, it is not a casual measure against VR sickness. In addition, the content that can be experienced is limited if the natural environment is reproduced in the VR environment.

The second is to suppress the occurrence of vection by blocking or adjusting the equilibrium information and peripheral visual field information related to spatial recognition.

The human eye can see a range of about 200° in the horizontal direction and about 125° in the vertical direction when the front is 0°, and this range is called the field of view.

A. The discriminative visual field is the central region (within about 5°) that has excellent visual functions such as visual acuity and color discrimination.
B. Effective visual field is an area that can be instantly receptive only by eye movement (horizontal about 30°, vertical within about 20°).

C. Stable gaze is the area where you can gaze comfortably with eyeball/head movement (horizontal 60°–90°, vertical 45°–70°, range where effective information reception is possible).
D. The guided visual field has low discriminating ability, but the area where the viewing angle information affects the spatial coordinate system (horizontal 30°–100°, vertical 20°–85° range has a large effect).
E. The auxiliary visual field (peripheral visual field) is a region where the presence of stimuli can be seen (horizontal 100°–200°, vertical 85°–130°) [7].

The central visual field of the human eye is better at identifying given information (color, shape, etc.) than the peripheral visual field. On the other hand, the peripheral visual field is good at sensing changes in the position and shape of an object, and performs motion perception [8]. Therefore, it has been clarified that the peripheral visual field is deeply related to the occurrence of vection, and that the addition of a mask area that covers the peripheral visual field makes it difficult for vection to occur [9].

4 Development of Virtual Horizontal Frame

A virtual horizontal frame created by Unity is displayed on the VR screen to reduce the occurrence of vection and confusion in the space recognition function, and to reduce VR sickness. The virtual horizontal frame was created with Adobe Photoshop CS6, and was read as a Sprite in Unity. It also performs rotation according to the head movement, programmed by a C # script.

4.1 Creating a Coaster

In the previous study, we created a roller coaster using Unity's asset Animated Steel Coaster to create content that causes VR sickness. In the previous course, in order to induce and amplify VR sickness, we created a roller coaster that reproduces yaw, pitch, and roll movements in a VR environment and makes visual information and spatial information from somatosensory inconsistent. Hereinafter, this content is referred to as coaster α (Fig. 1).

We decided that a roller coaster, which is a mixture of fast linear motion and rotational motion, is the most suitable content for generating effective spins on the subject. In addition, since the roller coaster has multiple rotations of the roll axis, pitch axis, and yaw axis, it is considered that the occurrence of vection can be expected. In addition, realistic snow mountain objects are arranged so as not to lose the sense of up and down. For the roller coaster course, we used Unity's asset Track_and_Rails and created a course of about 1 min 30 s per lap. Hereinafter, this content is referred to as coaster β (Fig. 2).

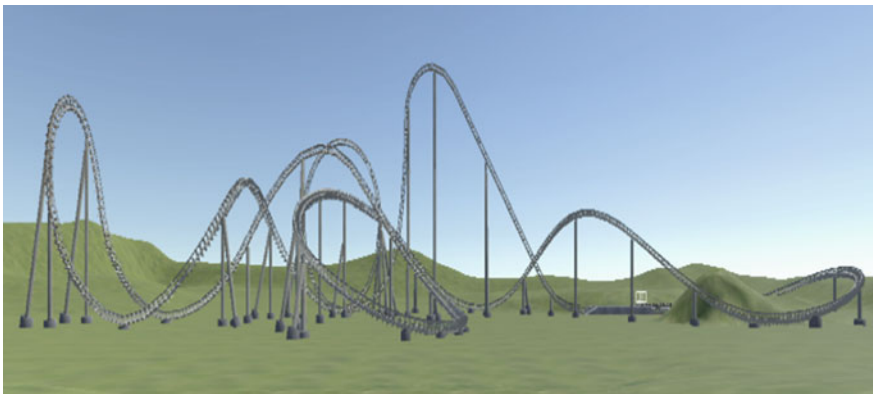


Fig. 1 Coaster α

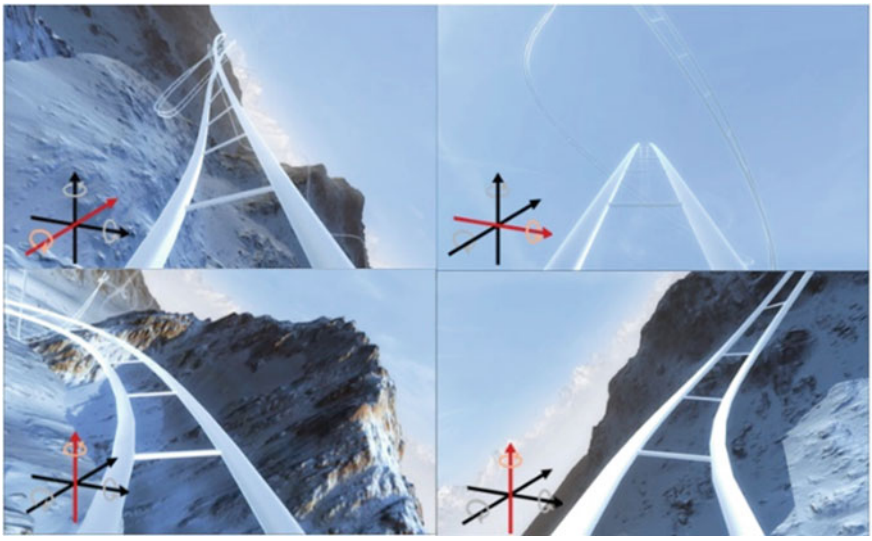


Fig. 2 Yaw, pitch, and roll axis course

We decided that a roller coaster, which is a mixture of fast linear motion and rotational motion, is the most suitable content for causing sickness, but since the viewpoint movement is constant, the target is placed on the course to increase the viewpoint movement. I decided to create a simple shooting game. In addition, by installing these, we tried to induce sickness more than a roller coaster. Using Unity's assets Tracks and Rails, I created a shooting course for about 1 min per lap, and created an original shooting game. In addition, by recording the score at the time of three laps, we improved the competitiveness and devised a way to concentrate more on the game (Fig. 3).

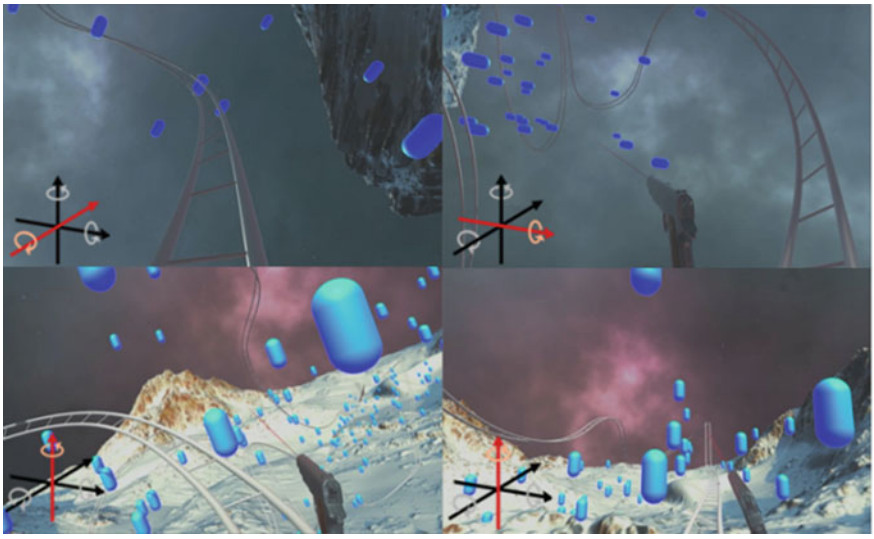


Fig. 3 Shooting coaster

Since the roller coaster moves automatically, we thought that content for the player to control the movement was necessary. We decided to create a game in which the movement method in VR is slide movement, because the player operates it and causes vection by linear motion and rotational motion more effectively. The outline of the game is to operate the avatar with the controller and collect the items placed in the map. In addition to stick operations, we also incorporated elements such as crouching buttons, jump buttons, and dash buttons to bring them closer to general game actions, and also tried to induce sickness when performing those actions. I created an original game using the map of RPG/FPS Game Assets for PC/Mobile of Unity assets (Fig. 4).

4.2 Virtual Horizontal Frame

In the previous research, we developed a frame that can easily grasp the angle of view on the horizon. The horizon part is not affected by the movement of the camera depending on the direction of the head, and is always displayed fixed on the screen. This makes it possible to always grasp the horizontal position of the field of view. The ring part rotates with the movement of the camera and gradually returns toward the horizon with a slow movement. This movement visually represents the horizontal position and the angle of inclination of oneself, and eliminates the gap in spatial recognition. In addition, the ring part hides the peripheral vision and works to suppress the occurrence of vection. However, since it is inevitable that displaying the virtual horizontal frame obstructs the view when viewing the VR content, the

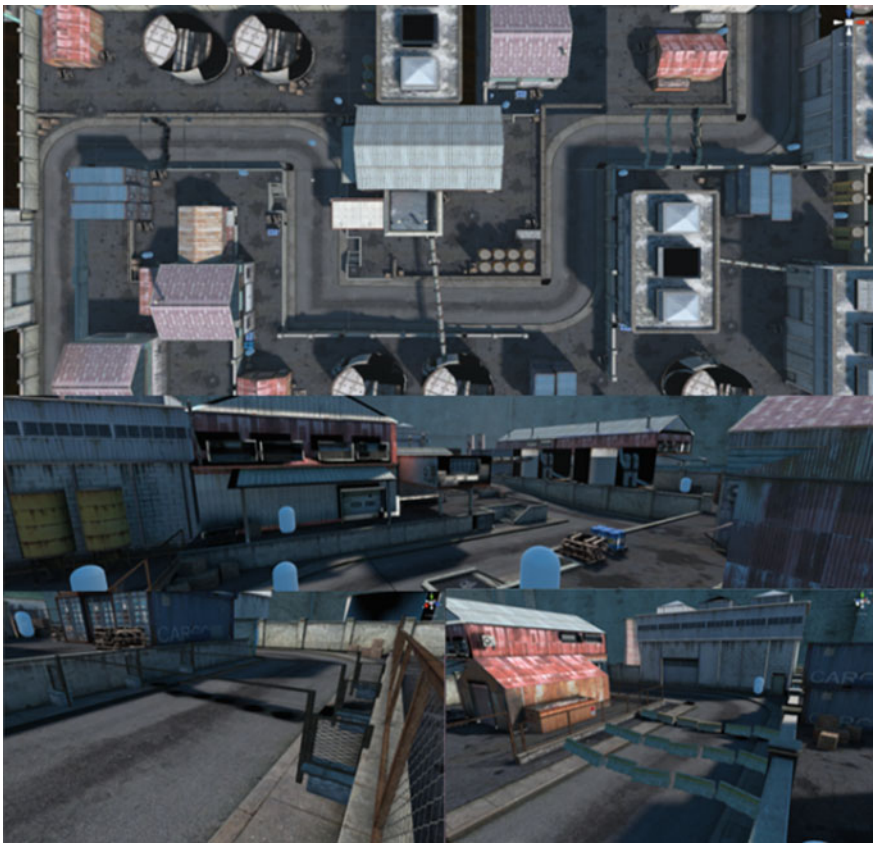


Fig. 4 Action game

range in which the VR content can be viewed is increased by using the broken lines for both the horizontal line and the ring, and the VR content is displayed. I try not to spoil the immersive feeling. Hereinafter, this frame is referred to as frame 1 (Fig. 5).

4.3 Omnidirectional Virtual Frame

In order to prove that the horizontal lines and rings of the virtual horizontal frame are effective in suppressing VR sickness, we created a virtual frame for comparison that obscures the peripheral visual field and provides a gaze point (Fig. 6).

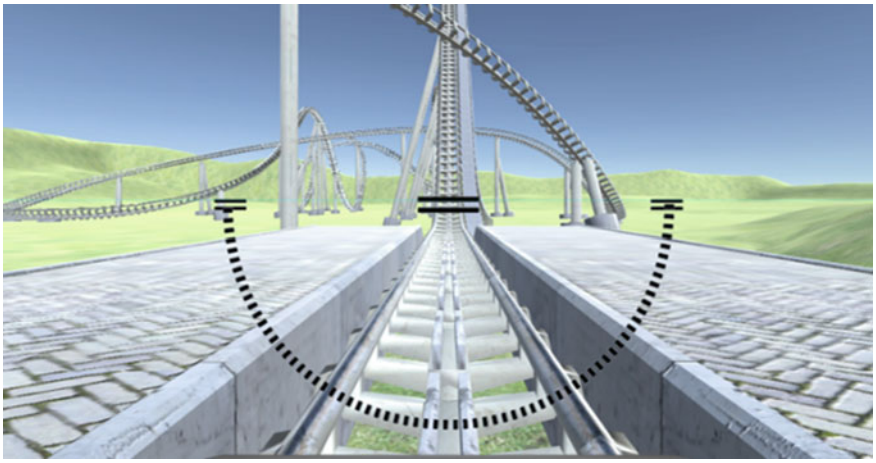


Fig. 5 Virtual horizontal frame

Fig. 6 Omnidirectional virtual frame

4.4 Horizontal Sensation Retention Frame with Limited Visual Field

We have developed a new virtual horizontal frame based on the virtual horizontal frame developed in the previous research. We have devised a frame design that inherits the rotation function of the horizon and ring of frame 1 and does not hinder the immersive feeling of the horizon and ring. In addition, the design of the horizon has been simplified. Therefore, it is assumed that the horizon is less likely to interfere with the content. In addition, in order to cover the design of the ring over the entire

Fig. 7 Horizontal sensation retention frame with limited visual field

peripheral visual field, the ring, which was a semicircle, was changed to a circle. It is speculated that this makes it possible to cover the peripheral visual field more and suppress vection. Also, by changing the gradation of the upper half of the circle to draw the horizontal line of the ring, the gradation and the boundary line of the lower half of the circle can be created. This boundary line is used as a new horizontal line. Hereinafter, this frame is referred to as frame 3 (Fig. 7).

The horizon part has been replaced with a black circle to provide a gaze point that does not interfere with the immersive feeling of VR content. This makes it difficult to grasp the horizontal position of the field of view, but its role as a gazing point does not change. Therefore, the effect of suppressing VR sickness on the horizon can be verified by comparing it with the virtual horizontal frame. In addition, by removing the rotation function of the ring part and changing from a broken line to a solid line, the frame is made to only cover the peripheral visual field. This makes it possible to verify how effective the rotation of the ring in the peripheral vision is in suppressing VR sickness. Hereinafter, this frame is referred to as frame 2.

5 Experiment

For the evaluation method of motion sickness, we used SSQ (Simulator Sickness Questionnaire) to perform subjective evaluation as a psychological index measurement. Although subjective evaluation is more prevalent in evaluation of motion sickness, VRsickness is thought to be caused by disturbance of autonomic nervous system due to sensory inconsistency of visual and brain perception.

5.1 Experimental Method

In order to evaluate VR sickness, SSQ and cerebral blood flow are measured. SSQ is used as a subjective evaluation. Cerebral blood flow is used as a physiological index. The reason for not using physiological indicators other than cerebral blood flow is that the study by Graybiel et al. Could not confirm a significant tendency in the onset of motion sickness and the tendency of fluctuations in physiological indicators [10]. On the other hand, Seraglia et al. Demonstrated that cerebral blood flow can be accurately measured even while experiencing VR in an experiment combining VR and cerebral blood flow meter (NIRS) measurement [11]. Therefore, this study verifies cerebral blood flow as a physiological index.

In this study, 5 subjects were assigned to each of the 4 contents, and a total of 20 male and female subjects cooperated. The subjects were asked to experience the content in each of the states of no frame, frame 1, frame 2, and frame 3, and then answered the SSQ questionnaire to compare the degree of sickness with and without each frame. In addition, when VR was attached, a cerebral blood flow meter was also attached at the same time, and changes in cerebral blood flow with and without each frame were compared. In addition to SSQ, we also conducted a simple questionnaire and decided to collect subjective evaluations for frames other than SSQ. In order to unify the measurement environment, the measurement was performed on a sunny day with more than 2 h after eating, and the room temperature was set to 25°.

5.2 SSQ Measurement

SSQ (Simulator Sickness Questionnaire) is one of the indicators of VR sickness. SSQ is a measurement of subjective evaluation, and was devised by Kennedy et al. In the United States as a method for diagnosing VR sickness using a drive simulator or an airplane simulator. Table 1 shows the 16 subjective evaluation items that are effective for simulator sickness by factor analysis of the subjective evaluation results obtained from a large number of simulator users. Ask them to answer one of the evaluations of each item from 0 to 3 (0 is no symptom, 3 is severe) and describe it. Those that correspond to nausea (N: Nausea), eyestrain (O: Oculomotor), and light-headedness (D: Disorientation) are marked with 1, and those that do not correspond are marked with 0. After describing the evaluation, each value is calculated by Eq. (1). In addition, a comprehensive index for VR sickness can be obtained by calculating the total value of nausea, eyestrain, and light-headedness (TS below Total Score) using Eq. (2) [12, 13].

$$\mathrm{N} = 9.54 \times \begin{pmatrix} \text{rating of 1} + \text{rating of 6} + \text{rating of 7} + \text{rating of 8} \\ + \text{rating of 9} + \text{rating of 15} + \text{rating of 16} \end{pmatrix} \quad (1)$$

Table 1 Subjective evaluation items of SSQ

(i) Signs and symptoms of sickness	Evaluation value x_i (0–3)	Degree of discomfort N_i	Degree of eye fatigue O_i	Degree of dizziness D_i
(1) General discomfort	–	1	1	0
(2) Feeling tired	–	0	1	0
(3) Headache degree	–	0	1	0
(4) Eyestrain	–	0	1	0
(5) The degree of difficulty in eye focusing	–	0	1	1
(6) Increased saliva	–	1	0	0
(7) Amount of sweating	–	1	0	0
(8) Degree of nausea	–	1	0	1
(9) The degree of inattention	–	1	1	0
(10) The degree of light-headedness	–	0	0	1
(11) Blurred eyes	–	0	1	1
(12) Dizziness (eye open)	–	0	0	1
(13) Dizziness (eye closed)	–	0	0	1
(14) Degree of loss of balance	–	0	0	1
(15) Stomach discomfort	–	1	0	0
(16) Belching	–	1	0	0

$$\mathrm{O} = 7.58 \times \left(\begin{array}{c} \text{rating of 1} + \text{rating of 2} + \text{rating of 3} + \text{rating of 4} \\ + \text{rating of 5} + \text{rating of 9} + \text{rating of 11} \end{array} \right)$$

$$\mathrm{D} = 13.92 \times \left(\begin{array}{c} \text{rating of 5} + \text{rating of 8} + \text{rating of 10} + \text{rating of 11} \\ + \text{rating of 12} + \text{rating of 13} + \text{rating of 14} \end{array} \right)$$

$$\mathrm{TS} = 3.74 \times (\text{total of 16 ratings of N, O, D}) \tag{2}$$

In this study, cerebral blood flow is used as a physiological index of VR sickness, and a cerebral blood flow meter (NIRS) is used to measure it. NIRS is a device

that non-invasively measures blood volume in the brain and visualizes changes in blood volume in the frontal lobe. Hemoglobin, a blood component, scatters light, and when oxygen is bound to it, the degree of scattering changes. Light absorption in the near-infrared wavelength range is caused by oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb), both of which have different absorption spectra. If the molar extinction coefficient of oxy-Hb and deoxy-Hb is known, the concentration change of oxy-Hb and deoxy-Hb can be calculated by measuring the change in absorbance at two or more wavelengths. A graph of cerebral hemoglobin can be displayed by the software in real time [14].

Seraglia et al. Combined VR and NIRS measurement to measure brain activity (parietal lobe/occipital lobe) when active on VR, and proved that cerebral blood flow can be accurately measured even while experiencing VR [11]. In addition, oxy-Hb increases by actively exercising by one's own will, such as voluntary movement, but oxy-Hb changes significantly in exercise that is not related to intention due to external influences can not be observed [11, 15]. From these, changes in oxy-Hb can be used as parameters that reflect brain activity.

The results of each SSQ questionnaire when experiencing content without a frame, when experiencing with frame 1, when experiencing with frame 2, and when experiencing with frame 3 are for each of the four contents. It is summarized in the graph. Each graph is the average value of N, O, D, and TS of all the subjects who experienced the content.

Graph 1 is the result of SSQ of coaster α. In frame 1, all the values of N, O, D, and TS were significantly reduced compared to those without a frame. In the case of frame 2, although it decreased, the value did not decrease from frame 1. In frame 3, the value of D decreased significantly, and the value of TS decreased from frame 2, but did not decrease from frame 1.

Graph 2 shows the SSQ results of coaster β. The values of D and TS decreased in all frames, and frame 2 decreased the most. In addition, the value of O decreased only in frame 2 and increased in frame 3. The value of N resulted in an increase in all frames.

Graph 3 is the result of SSQ of the shooting game. The overall value of frame 1 decreased compared to that without a frame, and the value of D in particular decreased significantly. In the case of frame 2, the values of N, D, and TS decreased, but O increased a little. As with frame 1, all values of frame 3 were reduced, and the result was that it was further reduced than that of frame 1.

Graph 4 shows the SSQ results of the action game. Compared with the case without a frame, the values of N, O, D, and TS were all reduced by adding a frame. Frame 2 had a small decrease compared to other frames. In frame 1 and frame 3, the values of N and D decreased most in frame 1, the value of O decreased most in frame 3, and the value of TS became similar.

In addition, in the four measured contents, all D values were large when there was no frame. However, by adding a frame, it was confirmed that the value of D decreased in all the contents. In particular, Figs. 8 and 11 in frame 1, Fig. 9 in frame 2, and Figs. 8 and 10 in frame 3 decreased significantly.

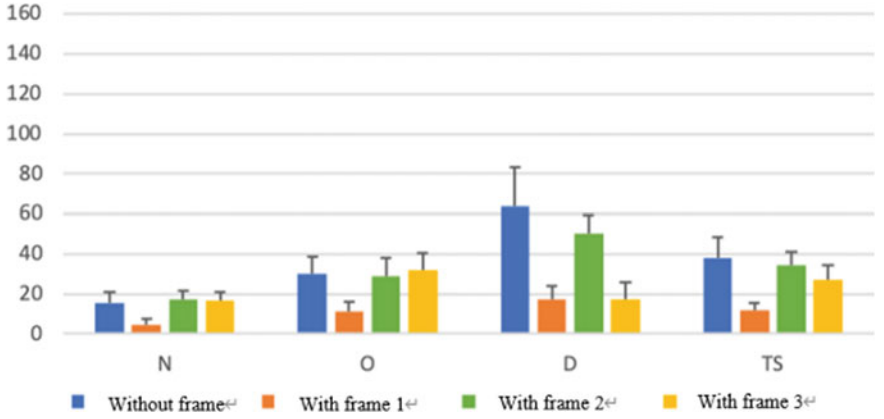


Fig. 8 [Coaster α] Average of each item of SSQ

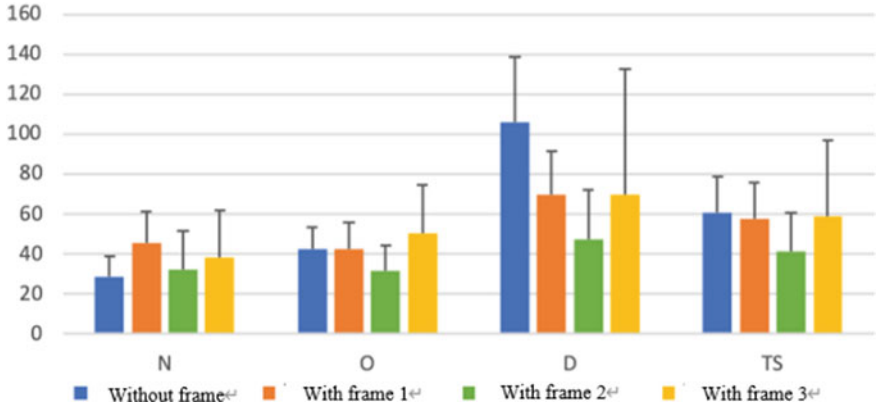


Fig. 9 [Coaster β] Average of each item of SSQ

From the above, the result is that the SSQ value is reduced overall for all frames compared to the case without frames. From this, it was confirmed that the system using virtual frames for various contents is effective in suppressing VR sickness. In addition, it was confirmed that frame 1 and frame 3 had lower SSQ values than frame 2, so the horizon is considered to be effective.

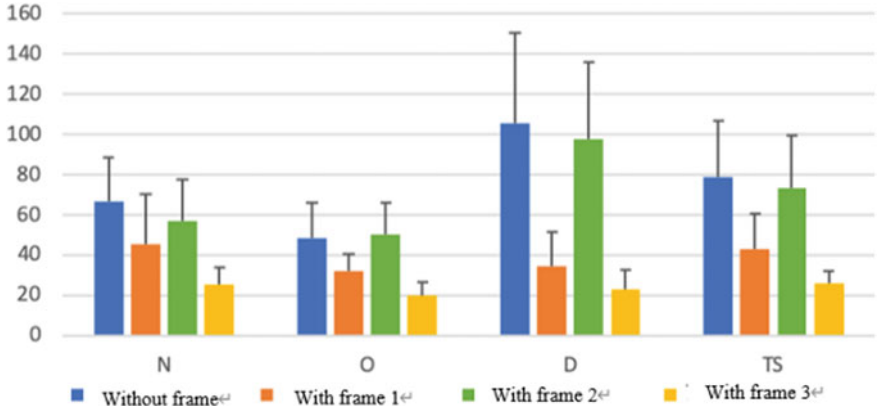


Fig. 10 [Shooting game] Average of each item of SSQ

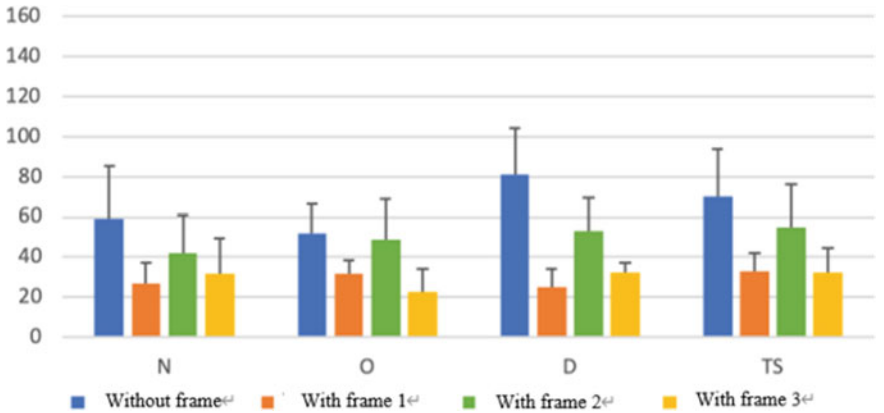


Fig. 11 [Action game] Average of each item of SSQ

6 Conclusion

In this research, we improved the virtual horizontal frame developed in the previous research and developed a horizontal sensation holding frame with visual field limitation. This suppresses VR sickness and enables a more immersive VR experience. From the experiment using the omnidirectional virtual frame developed on the way, it was found that the effect of suppressing VR sickness like the virtual horizontal frame cannot be obtained only by covering the peripheral visual field. In other words, the horizontal lines and rings of the virtual horizontal frame were greatly involved in suppressing VR sickness. In addition, in experiments using VR content, which is performed while the player is stationary and arbitrarily moves in 3D space by operating the controller, or while moving the body, the SSQ score was lowered after

experiencing each content displays a virtual frame. In other words, VR sickness can be suppressed by using a virtual frame. In addition, since the SSQ score of all contents has decreased, the virtual frame supports various contents.

When comparing the effect of suppressing VR sickness between the virtual horizontal frame and the horizontal sensation holding frame with visual field limitation, both were able to significantly suppress sickness, but no significant difference was observed overall. However, the newly designed frame that does not hinder the immersive feeling is more practical than other frames because it has the same effect of suppressing VR sickness as the virtual horizontal frame. In addition to this, we can formulate different visual specialization strategies according to different results of each SSQ questionnaire when experiencing content without a frame, when experiencing with frame 1, when experiencing with frame 2, and when experiencing with frame 3. Extend on this basis, we can also alleviate a certain symptom of VR sickness. We can even do the opposite, use visual elements to deepen the senses. Through seven experiments, there is an evidence for a speed-abstraction effect, where the perception of moving faster (vs. slower) leads people to rely on more abstract (vs. concrete) mental representations during decision making [16]. Adjust the human body sense through VR visual elements to generate a sense of speed that has a different sense of movement. In this way, the VR experience can be increased in a more three-dimensional manner. Even expand into more new directions of use.

References

1. Nobuhisa, T.: A survey of countermeasure design for virtual reality sickness. Trans. Virtual Real. Soc. Jpn. **10**(1), 129–138 (2005)
2. Nobuhisa, T., Takagi, H.: Design system of virtual reality environment based on VR presence and VR sickness. Trans. Virtual Real. Soc. Jpn. **11**(2), 301–312 (2006)
3. Reason, J.T., Brand, J.J.: Motion Sickness. Academic Press Inc. (1975)
4. Money, K.E.: Motion sickness. Am. Physiol. Soc. **50**(1), 1–39 (1970)
5. Kitagawa, E., Tanaka, S., Abiko, S., Tsukada, Y., Shiomi, K.: Method for reducing kinetosis using viewing mobile media in moving vehicles. J. Inst. Image Inf. Telev. Eng. **67**(11), J388–J399 (2013)
6. Sawada, Y., Itaguchi, Y., Hayashi, M., et al.: Effects of synchronized engine sound and vibration presentation on visually induced motion sickness. Sci. Rep. **10**, 7553 (2020)
7. Hatada, T. (Tokyo institute of Polytechnics): Measurement of information receiving and visual field. Jpn. J. Ergon. **29**, 86–88 (1993)
8. Kishishita, N., Orlosky, J., Kiyokawa, K., Mashita, T., Takemura, H.: Investigation on the peripheral visual field for information display with wide-view see-through HMDs. Trans. Virtual Real. Soc. Jpn. **19**(2), 121–130 (2014)
9. Ryu, J., Hashimoto, N., Sato, M.: Analysis of vection using body sway in immersive virtual environment. Technical report of IEICE (Multimed. Virtual Environ. **103**, 63–68) (2003)
10. Graybiel, A., Lackner, J.R.: Evaluation of the relationship between motion sickness symptomatology and blood pressure, heart rate, and body temperature. Aviat. Space Environ. Med. **51**, 211–214 (1980)
11. Seraglia, B., Gamberini, L., Priftis, K., Scatturin, P., Martinelli, M., Cutini, S.: An exploratory fNIRS study with immersive virtual reality: a new method for technical implementation. Front. Hum. Neurosci. **5**, 176 (2011)

12. Ujike, H.: Visually induced motion sickness. J. Inst. Image Inf. Telev. Eng. **61**(8), 1124 (2007)
13. Kennedy, R.S., Lane, N.E., Berbaum, K.S.: Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. Int. J. Aviat. Psychol. **3**, 203–220 (1993)
14. Shimadzu: Functional brain imaging (fNIRS). https://www.an.shimadzu.co.jp/apl/lifescience/invivo.html. Accessed 24 Oct 2020
15. Hirayama, K., Watanuki, K., Kaede, K.: Brain activation analysis of voluntary movement and passive movement using near-infrared spectroscopy. Trans. Jpn. Soc. Mech. Eng. Ser. C **78**(795), 3803–3811
16. Shani-Feinstein, Y., Kyung, E.J., Goldenberg, J.: Moving, fast or slow: how perceived speed influences mental representation and decision making. J. Consum. Res. (2022)

Represent Score as the Measurement of User Influence on Twitter

Yuto Noji, Ryotaro Okada, and Takafumi Nakanishi

Abstract In this paper, we present a novel method for quantifying a user's ability to spread information based on the number of Retweets (RTs) and Likes they receive on Twitter. In today's social network services, there are numerous users with the ability to spread information, called "influencers". However, even if they post the same content, the reactions they receive vary from user to user. Therefore, it is useful to create an index that represents the diffusion ability of each account to analyze diffusion behavior in social network services. In general, the diffusion status of information on Twitter is often quantified in terms of the number of RTs, Likes, and impressions of tweets alone. In this novel method, we propose a method for extracting indicators that show the diffusion power, not of tweets alone, but users as a unit, by measuring the ratio of the number of RTs and Likes based on the number of RTs and Likes of users in the past. The index obtained by this method can be used as an indicator for analyzing diffusion behavior on Twitter and may help conduct a more granular analysis. In this study, we conducted an experiment in which we collected tweets from 10 international celebrities for three years, divided them into multiple time series types, and applied this method to qualitatively evaluate them from the tweet text. The results showed that a bias exists when the period covered by the method is narrow, but when measured over a periodic unit of one year, there was no significant blurring, and it was possible to determine the status of the user in terms of the tweet text. We also found that each field was coherent and that there was a nature to the field.

Keywords Influence · Social network analysis · Twitter

Y. Noji · R. Okada · T. Nakanishi (✉)
The Tokyo Foundation for Policy Research, Department of Data Science, Musashino University, Tokyo, Japan
e-mail: takafumi.nakanishi@ds.musashino-u.ac.jp

Y. Noji
e-mail: s1922026@stu.musashino-u.ac.jp

R. Okada
e-mail: ryotaro.okada@ds.musashino-u.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_3

1 Introduction

Social networking services such as Twitter, Instagram, and Facebook have become important communication, information transmission, and information-gathering tools in the modern age. These services have replaced television and newspapers. They are used to obtain important information promptly, as well as to share daily life among friends. A common feature of these services is the ability to "share" posts. Recently, there are "influencers" who can information on these services as much as or more than television personalities and news organizations. Their posts are shared more than the typical user's. This means that the diffusion of information depends not only on the information itself but also on the one disseminating it. Many social networking services allow users to "Like" a post, which does not spread the information, but shows a liking for the post. Although text mining and machine learning have been used to study the diffusion of posts, we believe that it is promising to include user characteristics that consider the different nature of these reactions as a factor.

In this paper, we propose a method to extract a represent score, which is a measure of a user's diffusion ability, from the number of RTs and Likes of a Twitter user.

The main features of this paper are as follows.

- Collected tweets from target users based on the defined rules, and constructed a tweet database including reaction information.
- We proposed a method to calculate the Represent Score, which is an index of diffusion power, using RTs and Likes in the tweet database. Represent Score shows the degree to which a user represents (speaks for) the ideas of other users.
- We also applied this method to the accounts of international celebrities and conducted a qualitative evaluation.

Most studies of user influence on Twitter use social graph analysis. These analyses mainly use in-degree (following), RT, mention and reply as features. Our study focuses on the ratio of RTs to Likes to determine the characteristics of user influence.

The remaining paper is organized as follows: In Sect. 2, we describe related research. In Sect. 3, we propose the Represent Score, a measure of diffusion power using the number of reactions to tweets on Twitter, which is our proposed method. In Sect. 4, we show summarizes the evaluation results and discussion, and Sect. 5 concludes the study and highlights future research.

2 Related Work

Our research focuses on user influence on Twitter. Therefore, this section presents the research on Twitter user influence. The Merriam-Webster dictionary defines influence as "the power or capacity of causing an effect in indirect or intangible ways." The role of influence and its effects is studied extensively in Sociology, Marketing, Communication and Political Science [1].

Numerous definitions of influence on Twitter have been proposed. Antonakaki et al. summarized the overall research on Twitter in a survey paper [2]. In the paper, the study of User Influence is positioned as a type of analytical study on Social Graphs. The mainstream method for calculating User Influence is to use PageRank [3] and its variants [4–9]; another method is to use betweenness centrality and its variants [10, 11].

Riquelme et al. presented 53 different Influence metrics on Twitter [12]. They summarized what features were used to calculate the various User Influence metrics. The features that can be obtained from Twitter are "follow-up relationships", "retweets", "mentions", "replies", and "favorite (like)".

Cha et al. conducted a study using large data sets on Twitter User influence [13]. They defined three metrics of User Influence: (1) Indegree influence (the number of people who follow the account), (2) retweet influence, and (3) mention influence. They investigated the characteristics of these metrics through a study using a large data set. They clarified that In-degree represents the popularity of a user; retweets represent the content value of one's tweets; mentions represent the name value of a user; the top users based on the three measures have little overlap.

In our proposed method, we focus on representativeness as one of the prominent indicators of User Influence, and we use the ratio of retweets to likes as a user characteristic. According to the paper [12], not many studies have focused on likes. This is probably because "like" does not create a new edge on networks, therefore has little influence in social graph studies.

We present two studies that use "like" in the definition of the metric for User Influence.

Hajian et al. defined a metric named Influence Rank (IR) for identifying opinion leaders on the Twitter social network [1]. They defined two types of social influence: one is informational social influence (the need to be right) and the other is normative social influence (the need to be liked). The IR for a node is defined as the average IR of its neighborhoods combined with another index called magnitude of influence.

Montangero et al. proposed a metric named TRank [14]. TRank is designed to find the most influential Twitter users on a given topic, defined through hashtags. This approach combines different Twitter signals and provides three different indicators that are intended to capture different aspects of being influent. The three indicators are as follows: followers influence (FI), retweet influence (RI) and favorite influence (FVI).

These two studies include "like" in their method of influence calculation based on graph networks. Our method uses only retweets and likes of a user and focuses on the user's characteristics rather than overall influence.

3 Proposed Method

In this section, we present our proposed method of extracting a represent score for measuring user influence on Twitter. This method collects users' tweets, aggregates their reactions to the tweets, and scores the users' diffusion ability based on the aggregate reactions. This is because the number of reactions on Twitter varies depending on the status of the influencer. The Represent Score is defined as the ratio of RTs and Likes as one of the indicators of the diffusion power of a post.

3.1 Overview

We show the overview of our proposed method in Fig. 1. The method consists of two main modules: Scraping, and Extracting Represent Score.

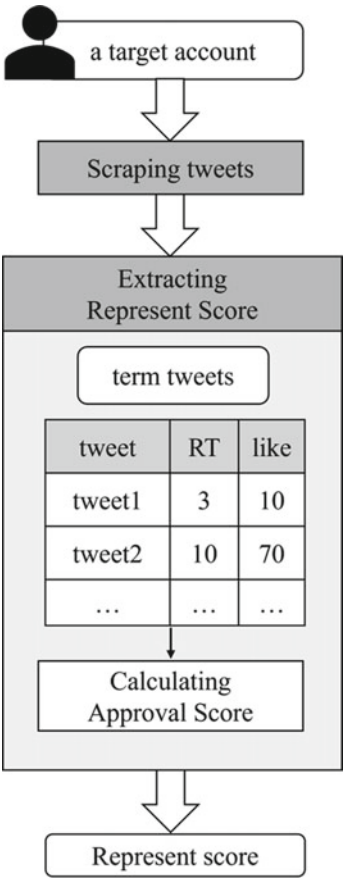


Fig. 1 Overview of proposed method. It consists of two main modules: scraping, extracting represent score, the user is specified, the user's tweets are collected, a tweet database is built, RTs and Likes are extracted from the reactions to those tweets, and the represent score is calculated based on those RTs and Likes

In our proposed method, we first input a username as a query and then collect tweets from the username by scraping. The number of likes and retweets are extracted from the collected tweets to build a database.

The Represent Score calculation for each account is applied to this database.

3.2 Scraping

In our method, we first enter a username as the search query, and then use the Twitter API or a tweet collection library to collect tweets from the target user. We search by specifying a period and collect all tweets during it. Replies are excluded during this period.

The reason for this is that replies are more of a response to other users, in other words, they are more conversational in nature. Since this method focuses on the ability of users to share their reactions to posts, replies that are not considered to be aimed for sharing by the influencer were excluded in consideration of this characteristic.

The collected tweets were saved as the "Database of tweets containing the keyword". In this database, we retrieve and save the information of "main text", "Like", and "Retweet" from the tweet.

3.3 Extract Represent Score

On Twitter, there are two types of responses to a tweet: "Like" and "Retweet". Like is a response that exhibits a positive feeling about the tweet. Retweet is a response corresponding to sharing a tweet.

In general, tweets receive more Likes than Retweets, but the ratio varies from account to account. We assume that the ratio of Likes to Retweets represents the characteristics of the account. For example, accounts that are rooted only in the offline community will have a smaller ratio of retweets. In contrast, accounts that talk about social issues will have a larger ratio of retweets. Even if two accounts have the same number of followers, they have different abilities to spread tweets. We defined represent score to express this propensity to be retweeted.

Twitter also has other reaction functions such as the number of replies and quoted retweets, but only the typical RT and Like functions have been used in this method to avoid complications.

The extraction represent score module consists of four steps:

1. Get the latest n tweets of the target account by scraping.
2. Extract the number of likes and retweets from those tweets.
3. Calculate the Approval Score with the following equation.

Represent Score is defined by the equation below.

$$A = \frac{\sum_i^n r_i}{\sum_i^n s_i}$$

A: approval Score
r_i: number of retweets of tweet i
s_i: number of likes of tweet i.

3.4 Select the Period of Time to Be Used in the Represent Score Calculation

Our method does not use all of the collected tweets in the database but selects the usable data by setting a time condition, relative to the date that the tweet was posted. This is to avoid bias differences due to environmental factors such as when reactions are more likely or less likely to occur depending on the period of interest and timing. For instance, this problem arises when comparing actors who are expected to get more reactions during the airing of their performances, artists whose reactions increase during the timing of their releases, or athletes whose competitions are held at specific times of the year. We believe that aligning the period of interest is an important factor in conducting an appropriately controlled experiment, except when applied to specific cases or topics with fixed timing.

4 Experiment

In this section, we conducted a validation to evaluate the method for calculation of the Represent Score. In Sect. 4.1, we describe the validation environment. In Sects. 4.2 and 4.3, we describe the time-series variation of the Represent Score. In Sect. 4.4, we prove the validity of our method by comparing qualitative tweets in addition to the comparison of Scores. In Sect. 4.5, we discuss the results of these validations.

4.1 Experiment Environment

In this experiment, we collected tweets from 2019 to 2021 from ten internationally active celebrities, two from each field. The fields were “politician”, “athlete”, “artist”, “actor” and “news”. A summary of target users and collected data is shown in Table 1.

There is a difference in the number of tweets, but no down-sampling was performed because it does not affect the calculation formula.

Table 1 Target information

Username	Type	Tweets count			Followers
		2019	2020	2021	
BarckObama	Politician	140	328	274	131.5 M
JoeBiden	Politician	1814	3304	486	33.7 M
BBCBreaking	News	897	704	721	49.7 M
cnnbreak	News	3802	4497	3191	62.8 M
10Ronaldinho	Athlete	155	83	160	20.5 M
Cristiano	Athlete	178	138	120	98.8 M
jtimberlake	Artist	55	67	32	114.3 M
justinbieber	Artist	134	528	203	63.2 M
selenagomez	Actor	57	237	174	65.7 M
Schwarzenegger	Actor	279	310	262	5.1 M

Figure 2 shows the distribution of RTs and Likes for each tweet for each user in 2021. As can be seen in the figure, the scale of reactions is coherent for each user. In this environment, there is a correlation between RTs and Likes, with a correlation coefficient of 0.915956.

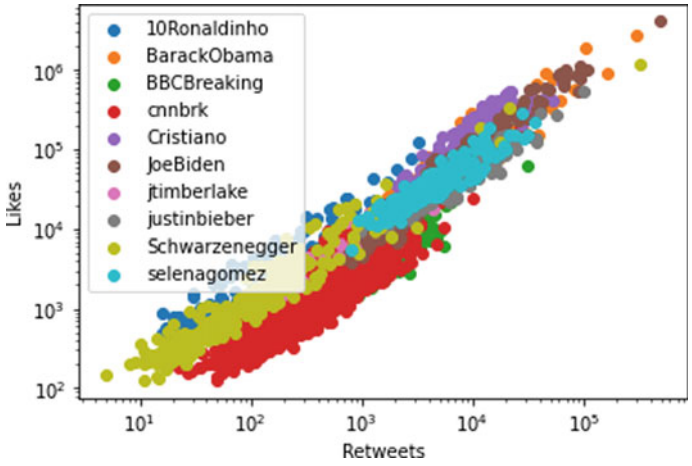


Fig. 2 RT and Like distribution chart. Scatterplot of RTs and Likes per tweet in 2021 for 10 subjects, displayed as a logarithmic graph

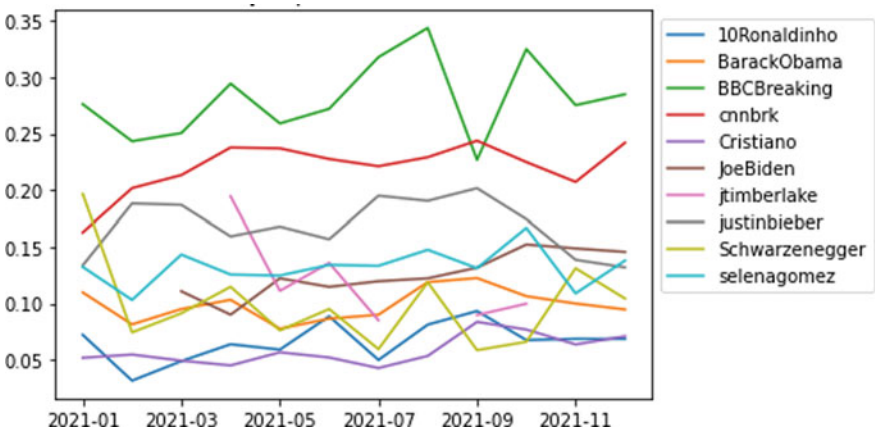


Fig. 3 Monthly represent scores in 2021. This figure shows the results of calculating the represent Score for the 10 target accounts' tweets in 2021 by tallying the reactions by month

4.2 Monthly Represent Scores

Figure 3 shows the monthly Represent Score in 2021. First, using the data presented in Sect. 4.1, we extracted the monthly Represent Score for each user. The results show a range of scores for the same user at different times of the year. As shown in Sect. 3.4, this indicates a bias depending on the time of year of the data used to calculate the score. This can also be taken as an indication of the score for the topic during that time period.

4.3 Yearly Represent Scores

Figure 4 shows a comparison of each user's Represent Score by year. The data presented in Sect. 4.1 was also used to extract each user's Represent Score for each year in the three-years: 2019, 2020, and 2021. The results show that, as with the monthly results, there is a range in the user's Represent Score depending on the time of year. However, the rank order of the scores did not change significantly, and the same trend can be seen even if the time period changes.

As with the monthly scores, there is a range of fluctuation, but because the scores were extracted over a cyclical period of one year, they are considered to represent changes in a person's status, rather than in a topic.

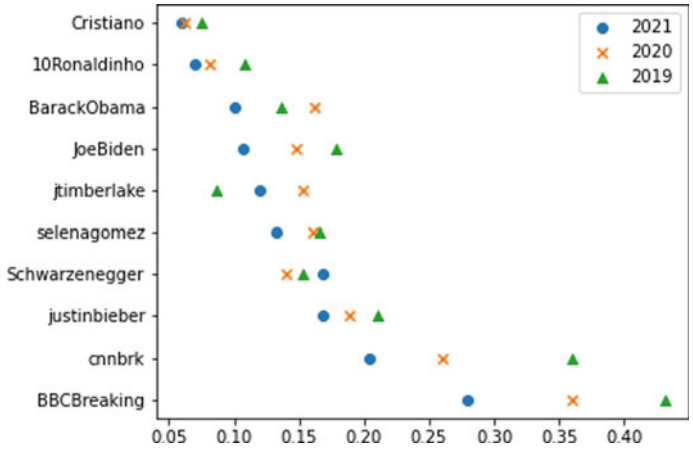


Fig. 4 Yearly represent scores in 2021. This figure shows the results of calculating the Represent Score for tweets from 2019 to 2021 for the 10 target accounts by tallying reactions by year

4.4 Most Retweeted Tweets Among the Collected Celebrity Tweets, Excluding Outliers

Table 2 shows the most retweeted tweets from the users shown in Sect. 4.1, excluding outliers; The Represent Score column lists each user's 2021 Represent Score, which is sorted in an ascending order. This shows that posts by users with high Represent Scores, such as "cnnbreak", "BBCBreaking", and "justinbieber" are intended to reach a large audience. For example, the first two accounts are news media accounts, which specialize in sending out news related information, and "justinbieber" is an artist who sends out information about new arrivals.

In addition, tweets that receive RTs from users with low Represent Score appear to be daily tweets or highly communicative rather than informative. For example, the tweet posted by "Cristiano" is topical, but lacks the importance of information, as it reports on his achievements. "BarackObama" is an accomplished and influential politician, but the post is directed to a specific individual and is long. Even if these users with low Represent Score have potential influence, the content of their tweets and the structure of their sentences suggest that they are less inclined to spread the information.

4.5 Discussion

In Sects. 4.2 and 4.3, depict the existence of bias depending on the time period used for score calculation. At the same time, the fact that the score fluctuates depending on the timing also indicates that the Represent Score can be characterized for detailed

Table 2 Most retweeted tweets. excluding outliers

Username	Followers	Represent score	Tweet	RT	Like
Cristiano	98.8 M	0.059133	Thank you to the Guiness World Records. Always good to be recognized as a world record breaker. Let's keep trying to set the numbers even higher!	32 K	455 K
10Ronaldinho	20.5 M	0.069529	Foi pra onde? Kkkkkk	2.4 K	34 K
BarackObama	131.5 M	0.100082	Happy anniversary, Miche! Over the past 29 years, I've loved watching the world get to know you not just as a daughter of the South Side, but as a mother, lawyer, executive, author, First Lady, and my best friend. I can't imagine life without you	18 K	341 K
JoeBiden	33.7 M	0.106535	We're seeing a coordinated attack on voting rights in this country. It's Jim Crow in the twenty-first century, and it must end.Congress must enact legislation to make it easier for all eligible Americans to access the ballot box and prevent attacks on the sacred right to vote	10 K	74 K
jtimberlake	63.2 M	0.120028	15 years ago today, I released my 2nd album… FutureSex/LoveSounds. This album changed my life. Every album is a different chapter and special to me but, this one?? I don't even know if I have the words	2.8 K	29 K
selenagomez	65.7 M	0.132857	Gracias a todos mis fans por tan lindas palabras sobre "De Una Vez". Los amo a todos! Thank you to all my fans for such kind words about "De Una Vez." I love you guys!	13 K	110 K

(continued)

Table 2 (continued)

Username	Followers	Represent score	Tweet	RT	Like
Schwarzenegger	5.1 M	0.168353	I hope all of our politicians stand on the side of the voters today. I'll be watching	0.5 K	7 K
justinbieber	114.3 M	0.168666	Justice World Tour 2022 International tickets on sale Friday	19 K	82 K
cnnbrake	62.8 M	0.204531	Police found multiple victims, including fatalities, at an office complex in Orange, California, after responding to a call of shots fired	0.7 K	15 K
BBCBreaking	49.7 M	0.279938	Duke and Duchess of Sussex tell Queen they will not be returning as working members of Royal Family	1709	9042

topics by appropriately narrowing down the time period and topics. If there is no topic or period condition in the analysis, it is considered better to calculate the Score by year to obtain the most recent stable Represent Score.

In Sects. 4.1 and 4.3, indicate that even though there is a strong correlation between RT and Like, the absolute difference in Represent Score is not very large, and it would not be possible to say that Represent Score alone shows a precise difference in diffusion ability.

Next, we compare the Represent Score including the number of followers and the scale of reactions. First, in both cases, there seems to be no direct relationship between the size of the Represent Score and the number of followers. This is probably because the Represent Score is a ratio index. However, further insight into the power of diffusion can be gained by examining the relationship between these three quantitative pieces of information among users. For example, compare "JoeBiden" and "BarackObama," they are both American politicians with similar attributes of global visibility and voice as presidential officeholders, their Represent Scores are 0.100082 and 0.106535, respectively, which are very close to each other. Figure 2 shows that their reactions are similar in scale. However, when comparing the number of followers, "JoeBiden" has about a quarter of the number of followers. In this case, "JoeBiden" has more spreading power than "BarackObama" because "JoeBiden" has more reactions despite of his small number of followers. From this point of view, the ratio of RT and Like is used as the score in this method because of the difference in their characteristics. However, to indicate diffusion power, it is thought to be more effective to set another index or weight that includes the relationship between

the number of followers and the scale of reactions. Depending on the number of followers and the size of the reaction, these could be used as weights.

The order of the scores also shows that celebrities are grouped by field, with athletes' tweets being more community- oriented and news accounts as information providers being more likely to be shared. Politicians also have lower scores and political tweets from the "Schwarzenegger" account have more Likes compared to RTs. This suggests that each field has its characteristics related to diffusion power.

From Sect. 4.4, it can be assumed that there is a difference in the Represent Score between informational and community-oriented tweets when looking at the qualitative tweets for each tweet. However, since there are differences in reactions to each tweet, it is necessary to consider the nature of the user's usual posting content and its influence on the frequency of such content. In addition, although these 10 accounts were targeted as international celebrities in this study, the fact that they were grouped by field suggests the need to conduct verification in an environment that facilitates qualitative comparisons by narrowing down the fields and increasing the number of people targeted to examine the aforementioned number of followers and scale of reactions. By doing so, we believe that we will be able to qualitatively examine the information that we were not able to explore in depth this time.

5 Conclusion

In this paper, we focus on the nature of Retweet and Like reactions on Twitter and present a method to calculate the diffusion power of influencers using these reactions. With this method, the diffusion power of each user can be quantified as an index called the "Represent Score".

In order to validate this method; we conducted a qualitative evaluation of its effectiveness by selecting international celebrities from multiple fields. The results showed that a bias exists when the period of application is narrow, but when measured on a periodic basis of one year, there was no significant blurring, and it was possible to determine the status of users in terms of the tweet text. It was also found that there was a cohesiveness in each field and that there was a nature to the field.

In the near future, we will apply this analysis after organizing other reaction information, such as the number of quoted retweets and replies; and account information, such as the number of followers and fields of a user's account, to apply more detailed information to the analysis, conducting controlled experiments, incorporating such information into the formula, or developing a formula that consists of such information. New indicators are being proposed. In addition, it would be necessary to compare and improve the formula, for example, by comparing statistics on the ratio of tweets per tweet, rather than the ratio of totals.

As a development of this study, we believe that qualitative analysis of Twitter diffusion behavior is possible by combining this method with text mining, such as the sentence structure of the tweet body and the words it contains.

Acknowledgements This research is a product of the research program of The Tokyo Foundation for Policy Research. We would like to thank Editage (www.editage.com) for English language editing.

References

1. Hajian, B., White, T.: Modelling influence in a social network: Metrics and evaluation. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 497–500. IEEE (2011)
2. Antonakaki, D., Fragopoulou, P., Ioannidis, S.: A survey of twitter research: data model, graph structure, sentiment analysis and attacks. Expert Syst. Appl. **164**, 114006 (2021). https://doi.org/10.1016/j.eswa.2020.114006
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998). http://dl.acm.org/citation.cfm?id=297810.297827
4. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network for a news media? In: Proceedings of the 19th International Conference on World Wide Web - WWW '10, p. 591. ACM Press, New York, USA (2010). http://dl.acm.org/citation.cfm?id=1772690.1772751
5. Said, A., Bowman, T.D., Abbasi, R.A., Aljohani, N.R., Hassan, S.-U., Nawaz, R.: Mining network-level properties of Twitter altmetrics data. Scientometrics **120**(1), 217–235 (2019)
6. Weng, J., Lim, E.-P., Jiang, J., He, Q.: Twitterrank. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10, p. 261. ACM Press, New York, USA (2010). http://dl.acm.org/citation.cfm?id=1718487.1718520
7. Priyanta, S., Trisna, I.P., Prayana, N.: Social network analysis of twitter to identify issuer of topic using pagerank. Int. J. Adv. Comput. Sci. Appl. **10**(1), 107–111 (2019)
8. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Proceedings of the 20th International Conference Companion on World Wide Web - WWW '11, p. 113. ACM Press, New York, USA (2011)
9. Hirsch, J.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U.S.A. **102**(46), 16569–16572 (2005)
10. Ediger, D., Jiang, K., Riedy, J., Bader, D.A., Corley, C., Massive social network analysis: Mining twitter for social good. In: 2010 39th International Conference on Parallel Processing, pp. 583–593. IEEE (2010). http://ieeexplore.ieee.org/document/5599247/
11. Laflin, P., Mantzaris, A.V., Ainley, F., Otley, A., Grindrod, P., Higham, D.J.: Discovering and validating influence in a dynamic online social network. Soc. Netw. Anal. Min. **3**(4), 1311–1323 (2013)
12. Riquelme, F., González-Cantergiani, P.: Measuring user influence on twitter: a survey. Inf. Process. Manag. **52**(5), 949–975 (2016)
13. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.: Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 4, no. 1, pp. 10–17 (2010)
14. Montangero, M., Furini, M.: Trank: ranking twitter users according to specific topics. In: 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), pp. 767–772. IEEE (2015)

The Experience of Developing a FAT File System Module in the Rust Programming Language

Shuichi Oikawa

Abstract The Linux kernel is a heart of the operating system, and provides the abstractions of hardware resources and manages them to be used efficiently and effectively by application programs. The operating system kernel has been programmed by the C programming language. It has been a major system programming language for a long time due to its efficiency and simplicity. Recently, there is a move towards introducing the Rust programming language to the Linux kernel for the type and memory safety; thus, it is meaningful to explore the possibility of utilizing Rust to program Linux kernel modules. This paper describes the experience of developing the FAT file system as a kernel module in Rust employing Rust for Linux as a basis of the development. We performed the experiments to measure its execution costs, and found that its performance is comparable with the original FAT file system written in C.

Keywords Systems software · Operating systems · Programming languages · Linux · Rust

1 Introduction

The Linux operating system is used by a wide range of systems. It is employed by personal mobile devices, such as smart phones and notebook PCs, servers in data centers, and also large scale supercomputers, which equip with millions of processors; thus, it is a part of the social infrastructure that constitutes information societies around the world. Since they are used mostly everywhere, its vulnerability is certainly a threat to societies. Therefore, its safety is very important.

The Linux kernel is a heart of the operating system. The kernel provides the abstractions of hardware resources and manages them to be used efficiently and

S. Oikawa (✉)
School of Industrial Technology, Advanced Institute of Industrial Technology, 1-10-40 Higashiooi, Shinagawa, Tokyo, Japan
e-mail: oikawa-shuichi@aiit.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_4

effectively by application programs. Since the kernel needs to interact with hardware directly, it is programmed in system programming languages, which provide the functionalities to cope with hardware, such as inline assembler, controlling data allocation on CPU registers, and data representations that fit hardware registers. The C programming language has been a major system programming language for a long time because of its efficiency and simplicity; thus, it has been used to program a number of operating system kernels including the Linux kernel.

Recently, there is a move towards introducing the Rust programming language to the Linux kernel for the type and memory safety [9]. The C language has been used as the programming language for the Linux kernel. While C is a popular system programming language, it is not a safe programming language. It is not strictly type safe since the C compiler successfully compiles programs that include undefined behavior while it checks the types of variables and functions. It is not memory safe since the C compiler does not check the memory boundary. Such unsafety features of C are the root cause of various vulnerability and defects.

The Rust programming language is a novel system programming language. It features type safety and memory safety while it can be used for the low-level programming that enables Rust programs to interact with hardware. Such features enable safe programming and help the reduction of vulnerability and defects. Therefore, it is significantly meaningful for the reduction of threats to societies to explore the possibility of utilizing Rust to program Linux kernel modules.

This paper describes the experience of developing the FAT file system as a kernel module written in the Rust programming language. We employ Rust for Linux [11] as a basis of the development. While Rust for Linux defines the interface that enables kernel modules written in Rust, it does not provide Rust interfaces required for a file system module. Thus, the implemented file system module in Rust interacts with the data structures defined in C. We discuss the impacts of such implementation in the aspects of type safety and memory safety. We performed the experiments to measure the execution costs of the programs that operate on the original FAT file system written in C and our Rust FAT file system in order to compare their performance. The experiment results showed that the performance of our Rust FAT file system is comparable with the original FAT file system.

The rest of this paper is organized as follows. Section 2 describes the overview of Rust for Linux. Section 3 describes the design and implementation of a FAT file system module in Rust. Section 4 shows the preliminary experiment results. Section 5 describes the related work. Finally, Sect. 6 concludes this paper.

2 Rust for Linux: Introducing Rust to Linux

Rust for Linux [11] is an effort to introduce the Rust programming language to the Linux kernel. First, this section introduce the overview of the Rust programming language. Second, Rust for Linux is briefly described. Then, the development of the Linux kernel with Rust modules is described.

2.1 The Rust Programming Language

The Rust programming language is a novel programming language. Its development was announced as a Mozilla project in 2010 [3] in order to develop a web browser engine [1]. Now, its development is actively conducted by the Rust Foundation. The Rust programming language features type safety and memory safety. It is type safe since the Rust compiler can successfully compiles only well-defined programs, which do not exhibit undefined behavior. It is memory safe since its compiler can successfully compiles only programs, in which memory pointers are used correctly to point to valid memory.

The Rust programming language is a system programming language [4]. The highest priority feature of system programming languages is performance [2]. The Rust compiler produces efficient executable code since it does not perform garbage collection, which is a fairly straightforward way to realize memory safety but can be a source of inefficient execution. Rust realizes memory safety by the ownership model, which manages a memory region, i.e. an object, by assigning the ownership of an object to a variable. When the variable becomes out of scope, the object assigned to the variable is freed. It is possible to gradually transfer the source code from the C language to the Rust language since their programs can be linked and execute together.

2.2 Rust for Linux

Rust for Linux attracted major attention when it became a part of Linux-next in March, 2021 [8], which is followed by the request for comments (RFC) for its development [9] and also the announcement from Google to support it for the Android platform [12]. Linux-next is the source tree that is intended for integration testing and is managed separately from the main line kernel. The integration into Linux-next is a major step toward the integration into the main line kernel while it is not always guaranteed and it often takes a long time especially for projects that include a large amount of work. Actually, the latest patch of Rust for Linux posted for review is already a fourth time but is still to be considered experimental [10].

Rust for Linux defines a platform to develop kernel modules in Rust. Since the main body of the Linux kernel is the same, the majority of the source code remains written in C. By taking advantage of the feature of Rust that its programs can be linked with C programs, kernel modules written in Rust is linked with the Linux kernel. Since the main functionality of the kernel remains in the main body of the kernel, Rust modules must interact with the main body written in C.

Rust for Linux provides numerous interfaces defined in Rust as *traits* in order to interact with the main body of the Linux kernel. A trait in Rust defines a group of functions that can be invoked for its type; thus, it defines the behavior of the type. The traits provided by Rust for Linux define the interface and behavior of the kernel

data structures by grouping the data structures and their supporting functions. Just as C modules utilize the interfaces to the kernel data structures, Rust modules utilize those traits defined by Rust for Linux.

At this moment, Rust for Linux is just an add on to the Linux kernel. It is possible to gradually rewrite the source code from C to Rust. The development of the Linux kernel is rapid, and its source code tree is always renewed. In other words, new source code is added while old unused code is removed. The Rust foundation is supported by major technology companies, such as Amazon, Meta, Google, and Microsoft, and they are also among active major Linux developers. Once Rust for Linux becomes a part of the main line kernel, it is possible that a number of Rust modules will replace the current C modules, and the majority of the kernel will be written in Rust.

2.3 Building the Linux Kernel with Rust Modules

Building Rust for Linux means building the Linux kernel with the Rust support and Rust modules. The Rust support option is found in the General setup menu of the Linux kernel configuration. The option is displayed only when a compatible version of the Rust compiler is found. Sample Rust modules can be compiled by selecting the Rust sample option in the Sample kernel code menu of the Kernel hacking menu. Figure 1 shows the display image of the General setup menu, in which the `Rust support` option is enabled.

After the configuration, the Linux kernel is build. Rust for Linux requires the LLVM C compiler (clang) along with the Rust compiler in order to correctly match the compiler infrastructure. While only using clang by invoking the make command with the CC=clang option is possible, using the whole LLVM tool chain is the most portable way by invoking the make command with the LLVM=1 option. In this way, the binary utility programs other than the C compiler are taken from the LLVM tool chain.

2.4 Developing Linux Kernel Modules in Rust

The development of Rust modules is self contained. Rust modules can be written only in Rust, and there is no need to write C programs in order to link them to the main body of the Linux kernel. The following shows example kernel modules written in C and Rust. First, a C kernel module definition example is described. A Rust kernel module definition example is described, next.

A simple C kernel module can be defined as shown in Fig. 2. It just prints out a message when it is loaded and unloaded. `module_init()` specifies the function, which is `helloworld_init()` for this example, invoked when a module is loaded and initialized. Similarly, `helloworld_exit()` specifies the function, which is `helloworld_exit()`, invoked when a module exits and is unloaded. The

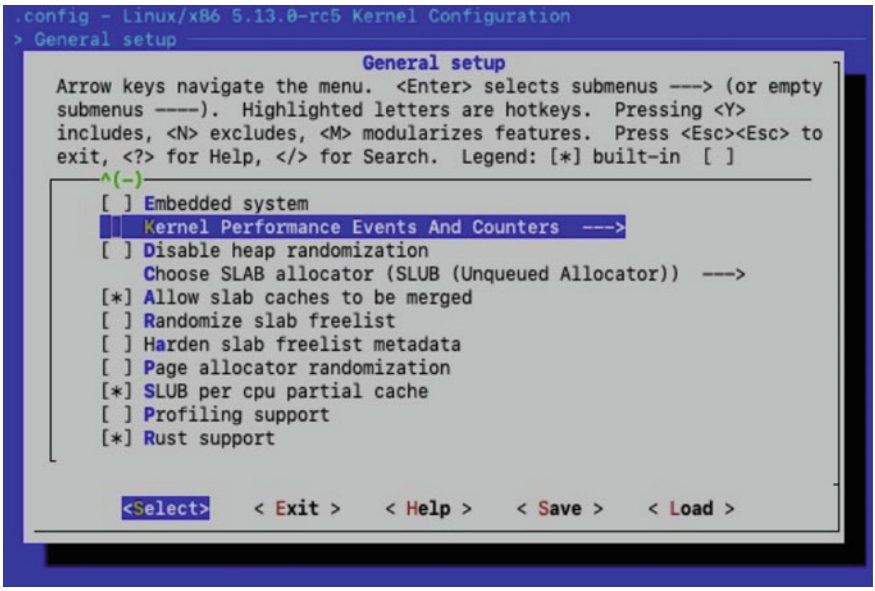


Fig. 1 Configuring Rust for Linux in the menuconfig command by selecting the Rust support option in the General setup menu

Fig. 2 A simple kernel module definition example in C


```
static int __init helloworld_init(void)
{
    printk("Hello World.\n");
}

static int __exit helloworld_exit(void)
{
    printk("Goodbye World.\n");
}

module_init(helloworld_init);
module_exit(helloworld_exit);
```

`printk()` function is similar to `printf()` and is used to print out a message in the kernel.

`module_init()` provides a mechanism to register a module with the kernel. Its argument is a pointer to the initialization function of a module, and is automatically placed in an array of a specific section. When the kernel is invoked and performs its initialization, kernel modules were initialized by calling their initialization function taken from the array. Similarly, `helloworld_exit()` registers the exit function.

A simple Rust kernel module can be defined as shown in Fig. 3. It behaves the same as the previous C example. `module!` is a macro description in Rust for Linux

Fig. 3 A simple kernel module definition example in Rust equivalent to the one in C

```
module! {
    type: HelloWorld,
    name: b"Hello World",
    author: b"Rust for Linux Contributors",
    description: b"hello world sample",
    license: b"GPL v2",
}

struct HelloWorld;

impl KernelModule for HelloWorld {
    fn init() -> Result<Self> {
        pr_info!("Hello World.\n");
        Ok(HelloWorld{})
    }
}

impl Drop for HelloWorld {
    fn drop(&mut self) {
        pr_info!("Goodbye World.\n");
    }
}
```

in order to register a module with the kernel. It replaces `module_init()` and `module_exit()` of a C module. `module!` specifies a module data structure by `type:`, which is `HelloWorld` in this example. `KernelModule` and `Drop` and the traits of the `HelloWorld`. The `init()` function of the `KernelModule` trait defines the module initialization function, and the `drop()` function of the `Drop` trait defines the module exit function.

3 Developing a FAT File System Module in Rust

This section first describes our motivation to develop a file system module in the Rust programming language. It then describes the interaction and interfaces with the kernel and the operations on the kernel data structures.

3.1 Motivation to Develop a File System Module

We developed a FAT file system module in Rust. A number of subsystems can be developed as modules in the Linux kernel, and a file system is one of such subsystems. Except for a few that require tight integration with the other subsystems, such as

procfs (process file system) and tmpfs (temporary file system), the file systems in the Linux kernel are implemented as kernel modules.

Rust for Linux provides several kernel modules as samples, and those modules can be used as references for implementation. Most of them are device drivers, and the rest of them are mere simple samples mostly similar to the Rust kernel module shown in Fig. 3. The status that there is no Rust kernel modules other than device drivers and mere samples motivated us to choose a file system as a target for development since a file system interacts with the kernel main body in a completely different way through different interfaces.

Next, the interaction of a file system with the kernel, the interfaces of the kernel used by a file system, and how they are realized in a Rust file system module are described.

3.2 Interaction and Interfaces with the Kernel

The interactions of a file system module with the kernel main body are typically performed in the following order:

1. Initialization of a module.
2. Registration of a file system.
3. Mounting a file system.
4. Mounting a block device.
5. Initialization of a file system.
6. Performing read/write operations.

Figure 4 illustrates the direction of the above interactions.

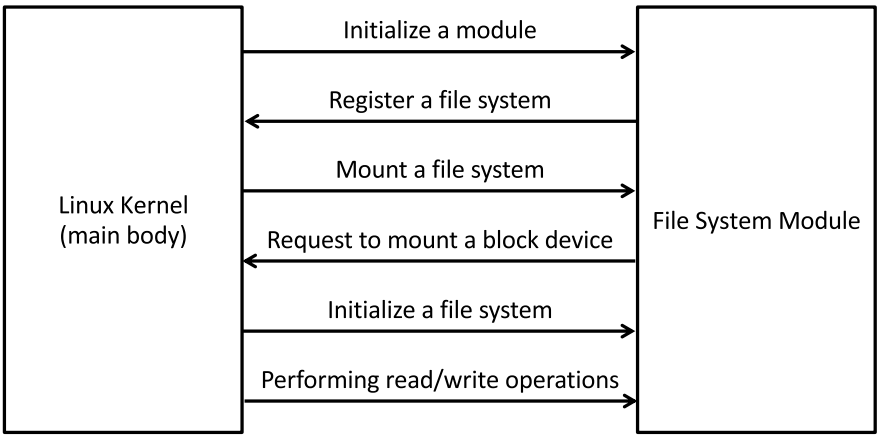


Fig. 4 The interactions of a file system module with the Linux kernel from the initialization to start read/write operations

The brief statement and the involving interfaces along with how each interaction is realized in Rust are described below.

(1) Initialization of a module is performed when the module is loaded into the kernel, and the main body of the kernel calls the specific initialization function of the module. The initialization function is registered to the kernel as shown in the Rust kernel module example of Fig. 3. The initialization performed in a module is different depending on the functionality of the module. The subsystems a module depends basically define the interfaces used for the initialization. If the module is a part of the file system subsystem, the module calls the interface to register it as a file system, and it is the second step described below.
(2) Registration of a file system is called by a file system module in order to literally register the module as a file system. Calling the `register_filesystem` interface registers a module as a file system with the kernel. The interface is in the main body of the kernel; thus, it is written in the C language. Since Rust for Linux does not provide the interface, its declaration in Rust was generated as shown in Fig. 5. The `register_filesystem` interface takes an argument of the `file_system_type` structure, which include the `mount` and `name` members. The `mount` member is a function pointer, and the `name` member is a character string that represents the type (or name) of a file system. If the file system type specified by the `mount` command matches the one specified by the `name` member of `file_system_type`, the main body of the kernel calls the function pointer specified by the `mount` member as the third step described below.
(3) Mounting a file system is called through the function pointer that was registered for the file system module as described above. Figure 6 shows the `msdos_mount` function that is called through the function pointer in order to mount the FAT file system. It is written in Rust for this work. The `msdos_mount` function needs to be callable from C programs since the `file_system_type` structure is defined in C and the main body of the kernel calls the `msdos_mount` function specified by its `mount` member. The `extern "C"` declaration is specified for this reason. Moreover, `unsafe` is specified since the function takes C pointers as its arguments. The function simply transfers the received arguments by calling the `mount_bdev` interface, which is the fourth step described below.
(4) Mounting a block device is performed by calling the `mount_bdev` interface. A file system defines a logical management structure in order to store data on a block device. Mounting a block device associates a file system with the block device

```
extern "C" {
    pub fn register_filesystem(
        arg1: *mut file_system_type
    ) -> c_types::c_int;
}
```

Fig. 5 The `register_filesystem` interface to register a module as a file system with the kernel

Fig. 6 The `msdos_mount` function called through the function pointer to mount the FAT file system

```
unsafe extern "C" fn msdos_mount(
    fs_type: *mut bindings::file_system_type,
    flags: c_types::c_int,
    dev_name: *const c_types::c_char,
    data: *mut c_types::c_void,
) -> *mut bindings::dentry {
    unsafe {
        bindings::mount_bdev(
            fs_type, flags, dev_name,
            data, Some(msdos_fill_super))
    }
}
```

and makes the file system its logical management structure. Figure 6 shows that the `mount_bdev` interface takes the `msdos_fill_super` function as its argument. It creates such an association between a file system and a block device through the initialization of a file system, which is the fifth step described below.

(5) Initialization of a file system is called through the `msdos_fill_super` function that is passed to the main body of the kernel by calling the `mount_bdev` interface as described above. Figure 7 shows the `msdos_fill_super` function that performs the initialization of the FAT file system. The function simply transfers the received arguments by calling the `fat_fill_super` function, which processes the detailed work of the FAT file system initialization. The initialization process first reads the management region (i.e. superblock) of the file system from the block device in order to obtain the file system management information, the location of the file allocation table, the block size, and so on, created on the block device. It then reads the root directory that is the start point to traverse to and perform read/write of directories and files on the file system, which is the sixth step described below.

(6) Performing read/write operations is done through the `inode_operations` interface, which is the structure of function pointers. The interface is created for directories and files, separately. Each contains the function pointers, that perform appropriate operations on them. While the `inode_operations` interface is defined in C, its functions are written in Rust for this work.

```
unsafe extern "C" fn msdos_fill_super(
    sb: *mut bindings::super_block,
    data: *mut c_types::c_void,
    silent: c_types::c_int,
) -> c_types::c_int {
    inode::fat_fill_super(sb, data, silent)
}
```

Fig. 7 The `msdos_fill_super` function performs the initialization of the FAT file system

3.3 Operations on the Kernel Data Structures

The Linux kernel defines various data structures that manage subsystems, abstractions, and data. The main body of the kernel passes those data structures to kernel modules as arguments. Kernel modules obtain necessary information from them, perform necessary processing, store the results in them, and share them with the main body of the kernel to return the results. Therefore, they play the important roles in order to realize the functionalities of the kernel.

The followings are the major data structures that file system modules operate on.

- `super_block`
- `dentry`
- `inode`
- `buffer_head`

The above data structures are defined for the following purposes. The `super_block` structure is for file system management information. The `dentry` structure is for a directory. The `inode` structure is for a file on storage. The `buffer_head` structure is for the read/write operations of data on storage.

Rust for Linux does provide no support for the above data structures as of the current date of writing this paper. It means that Rust modules directly operate on the data structures defined in C. In this case, Rust modules receive the raw pointers to the data structures as their arguments. Such operations impact the safety features provided by Rust. The impacts are discussed in the aspects of type safety and memory safety below.

Rust can guarantee type safety even for the data structures defined in C. They can be translated into the definitions in Rust automatically by the tool named *bindgen*, and the data structures can be transparently handled in C and Rust programs. Therefore, it is possible to write programs in C and Rust that process the same data of a certain type of a data structure. In terms of type safety that is not provided by C, Rust performs no implicit type translation. It requires explicit type translation everywhere necessary, and there are a number of places where explicit type translation is necessary in the Rust FAT file system module. It exposes the fact that the types used in the kernel source code are not cleanly defined. While it is cumbersome to write explicit type translation, it makes the type of variable and functions clear. Therefore, explicit type translation is effective to improve the readability of the source code.

Rust in principle does not operate on raw pointers for memory safety since there is no way to perform necessary checking at compile time. Rust requires the code to deal with raw pointers to be placed in the unsafe block in order to tell the Rust compiler that raw pointers are allowed in exchange for risk; thus, Rust does not provide memory safety for the code in the unsafe block. Rust modules inherently receive the raw pointers from the main body of the kernel. The raw pointers point to the data structures defined in C. Therefore, the code that interacts with the kernel interfaces ends up with being in the unsafe block.

4 Experiments

We performed the experiments to measure the execution costs of the programs that operate on the original FAT file system written in C and our Rust FAT file system in order to compare their performance. We first describe the experiment environment that are used for the experiments, and then show the measurement results.

4.1 Experiment Environment

We employ the QEMU system emulator as a target environment for experiments. The version of QEMU used for the evaluation is Sect. 2.1, and QEMU emulates x86_64. The configuration of QEMU used for the evaluation is 1 CPU, 256MB of RAM, and the virtio-blk paravirtual block device. Evaluation programs were stored in the initial ram disk.

The evaluation of execution costs needs to measure execution times. Times counted by the interrupts from a timer device are not accurate enough on system emulators. Instead, the number of instructions executed is used as the measure of execution costs. The `-icount 0` option of QEMU lets the TSC (time stamp counter) register count the number of instructions executed. The RDTSC instruction reads the value of TSC.

4.2 Measurement Results

We measured the execution times to read data from a file since it is the most basic operation of a file system. We created files, of which sizes are 1, 4096, and 8192 bytes, on the FAT file system, and measured the execution times to read data from them. The page cache is discarded before each measurement in order to read data from storage. The measurements were performed 1000 times, and their averages are shown as the results.

Table 1 shows the results of the measurements, and Fig. 8 depicts them. For a file of which size is just 1 byte, the performance of the Rust FAT file system is comparable with the original one. When the sizes of file are larger, however, the performance of

Table 1 The number of instructions to read data from a file

File size (Byte)	Rust FAT	Original FAT	Ratio (%)
1	32,579	31,245	104.3
4096	77,660	40,461	191.9
8192	79,800	55,690	143.3

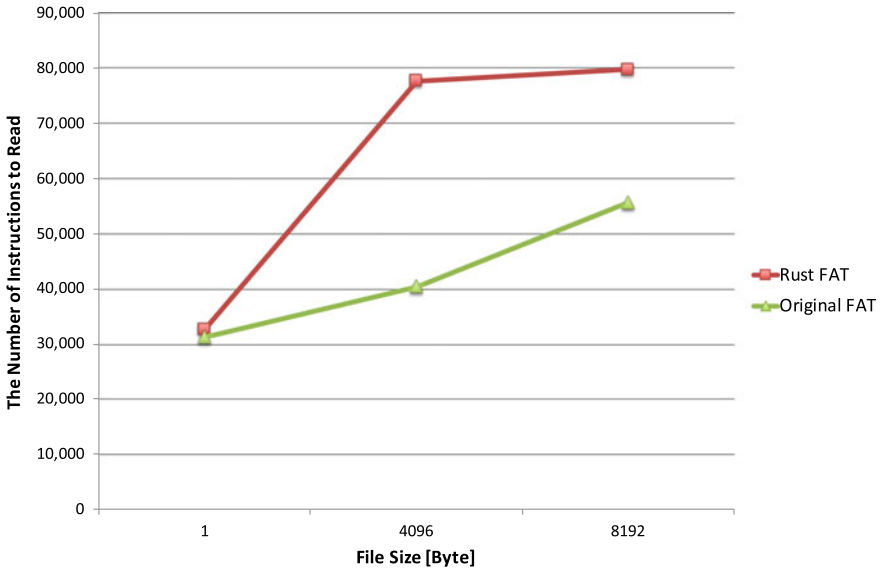


Fig. 8 Comparison of the number of instructions to read data from a file

the Rust FAT file system degrades significantly. The result does not match intuition. Since a file system locates the block number of data in storage and the main body of the kernel performs data transfer, the cost to read a 1 byte file is the basic cost mostly affected by a file system. We will investigate the reason of the results.

5 Related Work

The language features of Rust as a system programming language are discussed in [2, 4]. Balasubramanian et al. [2] discusses the mechanisms that were previously hindered by the high cost and complexity of their implementation but can be enabled by the capability of Rust. Jung et al. [4] discusses the requirements of systems programming, of which main concern is the control of performance and resource constraints. It argues for the features of Rust that are favorable for the requirements. While they focuses on the language features, this paper showed the implementation of a realistic file system module and discussed the impacts explored through the implementation in the aspects of type safety and memory safety.

There are several studies that work on the implementation of the kernel and a file system in Rust. Levy et al. [5] describes the implementation of an embedded kernel in Rust. The kernel features only a small set of mechanisms that deal with threads, interrupt and exception handling, and I/O memory mapping. RedLeaf [7] is a new operating system that aims to show that the type and memory safety features

of Rust can enable language-based isolation domains. It is based on the microkernel architecture, and it realizes the simple POSIX subsystem in an isolation domain. While they leverages the features of Rust in order to explore new abstractions and mechanisms, it takes time for them to be employed in practical systems. Bento [6] is a file system implementation in Rust. Since it is implemented as a user-level service by using the FUSE interface, it is different from our work that is implemented as a kernel module.

6 Summary

Recently, there is a move to introduce the Rust programming language to the Linux kernel for the type and memory safety. This paper described the experience of developing a FAT file system module in Rust employing Rust for Linux as a basis of the development. We implemented the FAT file system as a kernel module in Rust. We performed the preliminary experiments to measure its execution costs. We found that they are comparable with the original FAT file system for a small file while the performance of the Rust FAT file system degrades significantly for larger files. Our future work include the investigation of the reason of the performance degradation and the completion of the file system development by implementing the missing functionalities. Moreover, the investigation of implementing the traits necessary for the interface with file system modules.

References

1. Anderson, B., Bergstrom, L., Goregaokar, M., Matthews, J., McAllister, K., Moffitt, J., Sapin, S.: Engineering the servo web browser engine using rust. In: Proceedings of the 38th International Conference on Software Engineering Companion, ICSE '16, pp. 81–89. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2889160.2889229
2. Balasubramanian, A., Baranowski, M.S., Burtsev, A., Panda, A., Rakamarić, Z., Ryzhyk, L.: System programming in rust: Beyond safety. In: Proceedings of the 16th Workshop on Hot Topics in Operating Systems, HotOS '17, pp. 156–161. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3102980.3103006
3. Hoare, G.: Project servo: technology from the past come to save the future from itself. http://venge.net/graydon/talks/intro-talk-2.pdf(2010). Accessed 01 July 2022
4. Jung, R., Jourdan, J.H., Krebbers, R., Dreyer, D.: Safe systems programming in rust. Commun. ACM **64**(4), 144–152 (2021). https://doi.org/10.1145/3418295
5. Levy, A., Campbell, B., Ghena, B., Pannuto, P., Dutta, P., Levis, P.: The case for writing a kernel in rust. In: Proceedings of the 8th Asia-Pacific Workshop on Systems, APSys '17. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3124680.3124717
6. Miller, S., Zhang, K., Chen, M., Jennings, R., Chen, A., Zhuo, D., Anderson, T.: High velocity kernel file systems with bento. In: 19th USENIX Conference on File and Storage Technolo-

gies (FAST 21), pp. 65–79. USENIX Association (2021). https://www.usenix.org/conference/fast21/presentation/miller
7. Narayanan, V., Huang, T., Detweiler, D., Appel, D., Li, Z., Zellweger, G., Burtsev, A.: RedLeaf: isolation and communication in a safe operating system. In: 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20), pp. 21–39. USENIX Association (2020). https://www.usenix.org/conference/osdi20/presentation/narayanan-vikram
8. Ojeda, M.: Linux-next rust support. https://git.kernel.org/pub/scm/linux/kernel/git/next/linux-next.git/commit/?id=c7c4c9b88eecb28d3b48ba855cd6f9f7391a33b2. Accessed 01 July 2022
9. Ojeda, M.: [rfc] rust support (2021). https://lkml.org/lkml/2021/4/14/1023. Accessed 01 July 2022
10. Ojeda, M.: [patch v4 00/20] rust support (2022). https://lkml.org/lkml/2022/2/12/123. Accessed 01 July 2022
11. Rust for linux. https://github.com/Rust-for-Linux/. Accessed 01 July 2022
12. Rust in the android platform. https://security.googleblog.com/2021/04/rust-in-android-platform.html. Accessed 01 July 2022

Factors Influencing Consumers' Online Grocery Shopping Under the New Normal

Satoshi Nakano

Abstract Adoption of online shopping for groceries has remained low compared to other retail domains but has accelerated due to the long-term impact of COVID-19. This study aims to assess the psychological factors that influence consumers' actual online grocery purchases under the new normal by combining purchase panel data and survey data. This study confirms that online grocery purchase amount is affected by traditional utilitarian channel choice factors including perceived risk, search cost, price-consciousness, and quality-consciousness. In addition, under the new normal, the results reveal that staying at home (especially during weekdays) and having more time leads consumers to purchase more online. Further, consumers with higher anxiety about COVID-19 are more likely to purchase online. The insights from this paper help retailers and marketers develop customer strategies.

Keywords Online grocery shopping · COVID-19 · Stay at home · Channel management · Psychographics

1 Introduction

The retail environment has changed dramatically with the spread of COVID-19 as consumers become accustomed to a new way of shopping—the "new normal" [31]. An increasing number of consumers is purchasing online, even groceries, for which online shopping adoption is still lower than in other online retail sectors. When consumers discover the convenience of e-commerce technology and this technology makes a significant change in their lives, new shopping styles will become habitual [32]. However, little is known about how different factors are affecting consumer shopping styles in the new normal than before the pandemic. Understanding this is important for marketers and retailers for designing their channel strategies.

For grocery shopping, studies have shown that many consumers do not purchase products online, but instead prefer to visit local offline stores [6, 24, 27]. This is

S. Nakano (✉)
Faculty of Economics, Meiji Gakuin University, Tokyo, Japan
e-mail: nakanost@eco.meijigakuin.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_5

because grocery shopping is based on more habitual needs and lower product differentiation compared to other product categories. Thus, making it difficult to take advantage of the online convenience [23, 24]. In addition, from the standpoint of logistics such as delivery and inventory management, it is difficult to take advantage of online retailing, especially when dealing with highly perishable products, for which a long-tail assortment may not be suitable [23] Hence, the adoption of online grocery shopping still lags behind other online retail domains. Driving consumers to this channel remains a challenge for retailers [11].

The motivation of this study is to assess the factors driving consumers' online grocery shopping under the new normal. Our research makes several contributions to the literature. First, we extend the research related to the determinants of consumers' channel choices. These determinants have been addressed by numerous studies. Using online channels, consumers gain the utilitarian benefits of convenience and time saving [7, 11, 24, 26], and product prices [9, 23], but they must also accept the disadvantage of purchase risks [17, 26]. It is also known that consumers gain hedonic benefits related to entertainment and exploration [21, 27]. However, it is unclear whether these factors contribute to online grocery shopping after COVID-19. In addition, since COVID-19 changed the way people work and spend their leisure time, we also need to pay attention to context-specific factors in consumers' lives of post-COVID-19. More and more people work from home on weekdays and spend their holidays at home as well. The perceived shopping values are also expected to change. Therefore, we consider context-specific factors in addition to the traditional channel choice factors.

Second, this study is among the first to provide an integrated view of consumer actual online shopping behavior and psychological factors after COVID-19. Most channel studies at the consumer or customer level can be divided into those based on behavioral data using scanner panel data or ID-POS data (e.g., [6, 7, 24]), and those using surveys (e.g., [11, 21]). The former is more accurate because they are based on actual behavior, but cannot consider consumers' internal factors. The latter can incorporate a variety of factors, but cannot evaluate whether those factors are linked to actual behaviors. To take advantage of both approaches, our study adopts an integrated approach by combining purchase panel data and survey data. In doing so, we aim to gain an integrated view of behavior and the psychology.

2 Theoretical Background

2.1 Online Grocery Retailing

Consumer adoption of online shopping has been the subject of numerous studies in multichannel and omnichannel retailing. It is a fairly mature research area [29]. Among these, the studies on consumers' online grocery shopping have also been ongoing since the early 2000s [9, 17, 23]. With the spread of the Internet and mobile, it

has been predicted that most consumers will adopt online grocery shopping. However, this is not necessarily the case at this time.

One provocative finding that emerges from a review of the literature [29] is the persistent existence of offline single channel consumers. Over the past two decades, retailers have faced the reality that consumers have been slower than expected to adopt online as a channel for purchasing groceries [11]. For example, in Japan, e-commerce (EC) for groceries has been steadily gaining popularity, but the EC rate is still low compared to other product categories such as electronics, clothes, stationery, and books. According to the Ministry of Economy, Trade and Industry [25], in 2019, the EC rate, which is the ratio of EC to total sales, for Japan's merchandise sales sector was 41.8% for office supplies and stationery, 34.2% for books, video and music software, 32.8% for electronics and 13.9% for clothing, while the rate for foods, beverages, and alcoholic beverages was only 2.9%.

There are several reasons why the EC is difficult to advance in the grocery sectors. The business model of grocery retailing has a low margin and high turnover. The degree of product differentiation among stores is small and leads to price competition [23]. In addition, the EC of grocery face challenges such as the heavy burden of picking and delivery and the limited capacity to accept orders. As a result, existing brick-and-mortar grocery retailers may continue to lose money when adding new online channels. Grocery retailers are finding it difficult to maintain both brick-and-mortar and online channels [11].

However, this situation could change. The pandemic has forced consumers to become more experienced online. This may well result in consumers who have remained stuck in the single channel migrating to the multichannel [29]. The question is how to predict the habitual formation of such behavior. In order to do that, it is important to update the factors that influence consumers' online grocery shopping in a new normal environment. In this study, we construct the behavior model based on the theory of consumer channel choice and the contextual factors after COVID-19.

2.2 Theory of Consumer Channel Choice

Previous research suggests that consumers' channel choice can be explained by the shopping utility maximization theory [6, 7, 34]. In line with this theory, consumers can be expected to select the channel that provides the highest overall acquisition and transaction utility. Acquisition utility refers to the net effect of the benefits and costs of products that need to be purchased. Consumers can receive benefits such as product quality, while they need to pay costs such as price. Transaction utility is the utility (or disutility) of delivering products from stores to homes. From the transactions of online shopping, consumers gain advantages such as convenience and time savings, while they accept disadvantages such as purchase risk. Online and physical stores each have positive or negative utility drivers, and the trade-offs between them determine consumers' channel usage.

Quality and price for acquisition. The key aspects of acquisition utility are quality and price [34]. In general, the advantage of online retailing is the "long tail," which allows retailers to deal with niche products. The depth and breadth of the assortment allow retailers to target customer segments with varying needs for product quality [30]. Quality-conscious consumers pay higher prices to get better product quality, and tend to be more loyal to their primary channel [5].

Price is a cost that consumers must pay, but online shopping can lead to savings benefits. Consumers have perceptions of pricing and value for each channel. When a sales price is lower than the product's perceived value, it is an incentive to use that channel [33]. Since consumers can easily access a wide range of information through the Internet, the online channel has the utilitarian benefit of price comparison ease. For many product categories, the online price tends to be lower than offline, but this is not always the case for groceries, where discounts may not apply [11]. This has to do with the low margins in grocery retailing and the cost of picking and delivery. Therefore, some research suggests that the role of price is less important in online grocery retailing [23]. However, the results are mixed. Conversely, [9] finds that consumers who shop online for groceries more frequently are more price sensitive.

Hedonic benefits (exploration and entertainment). In addition to these utilitarian benefits and costs, consumers can also obtain hedonic benefits when acquiring products. The exploration benefit is one of the hedonic benefits. By searching several channels, consumers get the opportunity to try new products. Consumers' propensity to seek hedonic exploration benefits is related to innovativeness [1, 21]. Reference [21] suggests that multichannel shoppers have a higher tendency to be innovative in their trialability of new and unpurchased products. Few grocery shoppers purchase everything online, and many of those who purchase online are multichannel shoppers who also use physical stores [23, 27]. In the grocery shopping context, [27] suggests that multichannel shoppers with more media touchpoints, such as individuals using multiple devices (e.g., PC and mobile phone) and social media, tend to be more innovative.

The entertainment benefit is also a hedonic benefit that consumers can gain from shopping [3]. The tendency to derive pleasure and excitement from shopping influences channel selection. Previous research reports the positive effects of shopping enjoyment on multichannel selection for search and purchase [21, 33].

Convenience and perceived risk for transaction. Transaction utility includes convenience and perceived risk. The online channel differs from offline grocery stores regarding the convenience of 24 h ordering capabilities and home delivery. Since online shopping allows consumers to make purchases without time constraints, the online channel provides consumers with time-saving benefits. Several studies suggest that there is a positive relationship between the lack of time and frequent online shopping [7, 24, 26].

Further, online channels provide convenience in the information search process. Using online channel, consumers have fast, optimal access to enormous amounts of product information. It has been proposed that the ease of searching for information online has a significant impact on the usage of online channels [33]. While online

purchasing has the benefit of convenience, consumers who have high loyalty to physical stores may not necessarily feel the convenience. Reference [21] shows that consumers who do not make online purchases and focus on stores tend to have higher store loyalty than multichannel shoppers.

Regarding transaction disadvantages, the perceived purchase risk is a major factor. This risk is higher for online purchases than for offline ones because consumers cannot have direct contact with the product or retailer. At times, the product the consumer receives may arrive damaged or may be different from what the consumer expected. In particular, a significant risk in online grocery is associated with receiving perishable food [26]. Previous research suggests that consumers who look to avoid such risks are less likely to make online purchases [17, 26]. However, some recent studies have shown that the perceived purchase risk does not strongly influence the adoption of online shopping. In the context of online grocery, [11] suggests that this is because consumers are more familiar and more trusting toward online shopping than a decade ago.

2.3 The Impact of COVID-19 on Consumers

COVID-19 has impacted consumer behavior worldwide in various ways. As several studies have shown, it has a tremendous impact on consumer channel usage in grocery purchases. Reference [12] discusses the impact of COVID-19 on the adoption of online shopping. Based on an online survey, they imply that the social share of confirmed COVID-19 cases increases the possibility of consumers purchasing food online. This tendency is also more pronounced among younger people, those who have a lower perceived risk of online purchase, and those who live in large cities. Reference [18] conducted a survey among U.S. households to evaluate utilization of online grocery shopping during the COVID-19 pandemic. According to their results, around 55% of respondents shopped for groceries online in June 2020, and of those, 20% were first-timers.

From a short-term perspective, context-specific factors of the pandemic related to health and stable living may influence consumers' online purchases. Reference [15] shows that illness and food shortage concerns increased the frequency of online shopping. Related to consumers' concerns, in the early stages of the pandemic, panic buying, which is the phenomenon of temporarily buying large quantities of foods and daily necessities, was seen in many parts of the world. Panic buying is caused by negative emotions such as anxiety and fear. The panic buying itself is a temporary phenomenon as the term "panic" implies. However, as the pandemic continues, consumers become more rational in their stockpiling and tend to hoard groceries on a regular basis [28]. In this study, we focus on the anxiety about COVID-19 to understand consumer behaviors during a pandemic. If the anxiety felt by consumers has a strong influence on online purchasing, the specific circumstances of the pandemic could be causing online purchases.

On the other hand, from a long-term perspective, we need to pay attention to consumers' living, as working styles are changing radically. For example, telecommuting may have reduced commuting time and increased leisure time. More people are spending more time at home, even on holidays. This could trigger online purchasing, but to best our knowledge, few studies have empirically demonstrated this. Hence, we also examine the impact of staying home status on online purchasing.

In addition, [32] raises the issue of changes in consume as consumers spend more and more time at home. While many things can be experienced from the comfort of home, [32] (p. 281) points out that "what we need is to empirically study 'IN-home everything' impacts consumer's impulse buying and planned vs unplanned consumption." Traditionally, it has been assumed that unplanned purchases occur mainly in physical stores. However, as online shopping increases, it is crucial to examine whether unplanned purchases also occur online. Therefore, we treat impulsiveness toward unplanned purchases as one of the contextual factors of new normal and examine whether this influences online purchasing.

2.4 Research Questions

This study focuses on grocery shopping, which has been slow to adopt online, and discusses how this is changing in the new normal. Figure 1 shows our conceptual model. We set the following research questions.

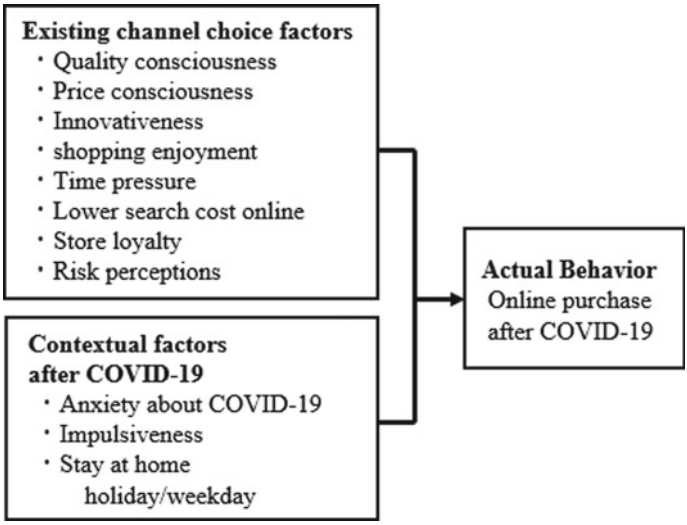


Fig. 1 Conceptual model

- RQ1. Which of the existing consumer channel choice factors influence online grocery purchase amount in the new normal environment?
- RQ2. What contextual factors related to after COVID-19 influence online grocery purchase amount?

Existing channel choice factors include quality consciousness, price consciousness, innovativeness, shopping enjoyment, time pressure, lower search cost online, store loyalty and risk perceptions. Contextual factors include anxiety about COVID-19, impulsiveness and the status of staying at home.

3 Methods

3.1 Data

We used a consumer scanner panel data, called the Syndicated Consumer Index (SCI), operated by the Japanese marketing research company INTAGE Inc. The SCI is the de facto standard scanner panel data used for marketing by many companies in Japan. The data document individuals' detailed purchase histories (purchase amount and unit). The product categories covered in this study are staple foods, seasonings, processed foods, snacks, ice cream, milk-based drinks, soft drinks, and alcohol (but not fresh fish, vegetables, or prepared box lunches). The data period is six months, from March 1 to August 31, 2020 from the beginning of the severe pandemic of COVID-19 in Japan until it settled down to some extent. From the data, the monthly online grocery purchase amount of each individual was extracted and used as the dependent variable for the analysis.

In addition, we conducted an online questionnaire survey between October 9, 2020 and October 19, 2020 among the panelists to collect their staying at home status and psychographics as described in detail later. Of the approximately 50,000 scanner panelists, we randomly conducted the online survey to the panelists. As a result, the online survey was delivered to 1,226 panelists and the valid responses were 883 (72.0%). These consumers are male and female between the ages of 20 and 69 who live in Japan.

The dependent variable, monthly online grocery purchase amount, includes a large number of zeros (i.e., a large number of consumers who did not make an online purchase during the month). This is a common phenomenon in data related to online shopping and has been observed in many other previous studies [2, 22]. To address this issue, we respond by using a Tobit regression model which corresponds the dependent variable containing many zeroes as described in the next section.

Moreover, actual purchase data such as scanner panel data often includes outliers on the right side. Therefore, we controlled for the effect of extreme outliers with reference to [22], which analyzed similar online purchase transactional data. We standardized the total online grocery purchase amounts for each consumer and dropped

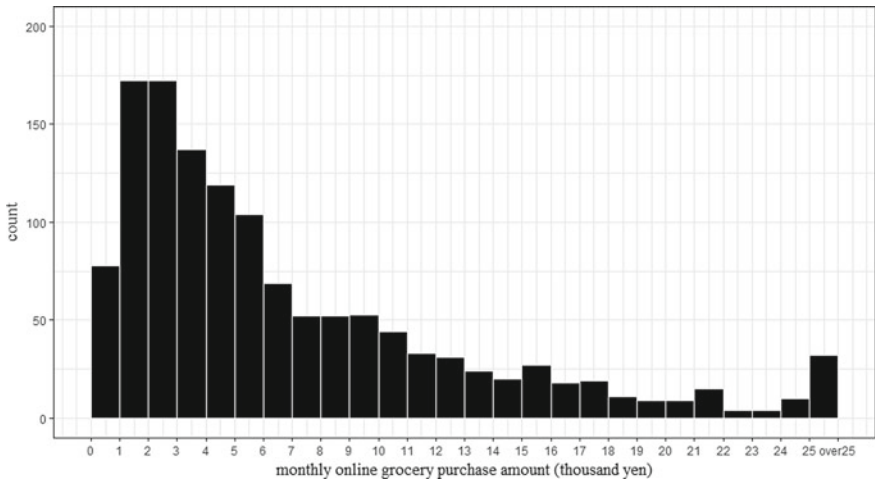


Fig. 2 Histogram of the monthly online grocery purchase amount (excluding 0)

consumers with standard scores of 4 or greater [16]. As a result, 12 consumers were excluded, resulting in a final sample of 871 consumers (total 5,226 opportunities for 6 months per consumer).

The summary statistics of the monthly online grocery purchase amount (thousand yen) are mean = 1.784, sd = 4.515, median = 0, the percentage of zero = 68.9%. Figure 2 shows the histogram of the monthly online grocery purchase amount (excluding 0).

3.2 Variable Measurements

We measured the psychographic variables using multiple items with five-point Likert scales ranging from 1 (fully disagree) to 5 (fully agree), as shown in Table 1. Reference [1]'s items were used for quality consciousness, price consciousness, innovativeness, shopping enjoyment, time pressure, store loyalty and impulsiveness. The constructs proposed by [1] have been used in many other previous studies for understanding consumer heterogeneity in a multichannel environment [21, 27]; thus, we also adopted them. For the lower search cost online, we used the items from [19]. For risk perceptions, we used [8]'s items. Anxiety was measured using four items following [35].

Confirmatory factor analysis was used to create psychographic variables (Tables 1 and 2). As a result of the analysis, the model fits the data well (GFI = 0.924, CFI = 0.946, RMSEA = 0.047). For reliability, Cronbach's α values for all constructs are greater than 0.7 [16]. The composite reliability (CR) values of all constructs

Table 1 Variable measurements

Construct with measurement items	Factor loadings
Quality consciousness	
It is important to me to buy high-quality products	0.707
I will not give up high quality for a lower price	0.705
I always buy the best	0.688
Price consciousness	
I find myself checking the prices even for small items	0.797
I compare prices of at least a few brands before I choose one	0.766
It is important to me to get the best price for the products I buy	0.745
Innovativeness	
I like to try new and different things	0.830
I am often among the first people to try a new product	0.768
When I see a product somewhat different from the usual, I check it out	0.735
Shopping enjoyment	
I like grocery shopping	0.831
I take my time when I shop	0.729
I enjoy grocery shopping	0.839
Time pressure	
I always seem to be in a hurry	0.951
Most days, I have no time to relax	0.844
I never seem to have enough time for the things I want to do	0.808
Lower search cost online	
It is cheaper to seek product information online	0.670
It does not take much to collect product information online	0.840
It is economical to search for information online before purchasing offline	0.694
Store loyalty	
I prefer to always shop at one grocery store	0.836
I am willing to make an effort to shop my favorite grocery store	0.893
Usually, I care a lot about which particular grocery store I shop at	0.840
Risk perceptions	
If I purchase online, there is a high possibility of getting the wrong product	0.966
It is difficult to judge the quality of a product online	0.858
It is likely that product I purchase online will not meet my requirements	0.707
Anxiety	
I feel tense about the COVID-19 pandemic	0.782

(continued)

Table 1 (continued)

Construct with measurement items	Factor loadings
I feel anxious about the COVID-19 pandemic	0.797
I feel stressed about the COVID-19 pandemic	0.675
I feel nervous about the COVID-19 pandemic	0.753
Impulsiveness	
I often find myself buying products on impulse in the grocery store	0.847
I often make an unplanned purchase when the urge strikes me	0.787

are greater than 0.7, which is above the cut-off criterion of 0.6 [4]. For convergent validity, average variance extracted (AVE) values of all constructs are greater than 0.5. For discriminant validity, we confirmed that the square root of the AVE for each construct is higher than its correlations with another construct [10].

Besides psychographic variables, we measured the status of staying at home on weekdays and holidays each month as time-variant variables. The question statement is: "Since the spread of COVID-19, I feel the need to stay at home on weekdays/holidays". The questionnaire was answered with a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The items are measured separately for weekdays and holidays on a monthly basis, from March to August 2020.

As control variables, we used demographics including gender, age, number of family members, income (million yen). Gender is expressed as a dummy variable, where 1 represents males. Other demographics are expressed as continuous variables. We show the summary statistics for staying at home variables and demographics in Table 3.

3.3 Model

We designed a Tobit model for panel data to predict the online purchase amount, denoted by y_{it}, which refers to consumer i's (i=1,…, N) online purchase amount in the month of year t (t=1,…, T). The Tobit model with individual random effect is used considering that the panel data formulation includes repeated observations for each individual. Since the distribution of online purchase amounts includes a large number of individuals who do not make any online purchases (equal to zero), we adopted a left-censored model. The Tobit model with a lower limit of zero can be described as follows:

$$y_{it}^{*} = \alpha + \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\gamma} + \mu_i + \upsilon_{it} \tag{1}$$

and

Table 2 Reliability, validity and correlation matrix

	α	CR	AVE	Correlation matrix									
				1	2	3	4	5	6	7	8	9	10
Quality consciousness	0.741	0.743	0.503	0.709									
Price consciousness	0.812	0.814	0.595	0.558	0.771								
Innovativeness	0.817	0.821	0.606	0.376	0.300	0.778							
Shopping enjoyment	0.839	0.841	0.639	0.455	0.425	0.396	0.800						
Time pressure	0.901	0.904	0.759	0.267	0.127	0.151	0.095	0.871					
Lower search cost online	0.738	0.752	0.51	0.414	0.224	0.156	0.303	0.153	0.714				
Store loyalty	0.891	0.893	0.736	−0.042	−0.089	−0.086	−0.113	0.184	−0.024	0.858			
Risk perceptions	0.875	0.883	0.719	0.068	−0.021	0.034	−0.002	0.166	0.037	0.483	0.858		
Anxiety	0.838	0.839	0.567	0.267	0.178	0.190	0.150	0.268	0.207	0.172	0.172	0.753	
Impulsiveness	0.799	0.799	0.665	0.115	−0.129	0.418	0.196	0.161	0.104	0.199	0.160	0.191	0.815

The diagonal values of the correlation matrix represent the square root of the AVE values. All other values represent the correlation coefficients

Table 3 Summary sample statistics

Demographics (N = 871)	%	
Gender (male = 1)	52.5%	
	Mean	SD
Age	49.061	11.558
Number of family members	2.799	1.232
Income (million yen)	5.939	2.905
The status of staying at home (N × T = 5,226 opportunities)		
Stay at home on weekdays	3.288	1.432
Stay at home on holidays	4.120	0.947

$$y_{it} = \begin{cases} y_{it}^{*} & if\ y_{it}^{*} > 0 \\ 0 & if\ y_{it}^{*} \leq 0 \end{cases} \tag{2}$$

where y_{it}^{*} refers to the unobserved latent variable. α is the constant term. μ_i is the random effect term that follows the normal distribution with mean 0 and variance σ_{μ}^{2}. υ_{it} is the remaining disturbance term which follows the normal distribution with mean 0 and variance σ_{v}^{2}. $\boldsymbol{x}_{it}$ is a vector of time-variant variables that capture the status of staying at home on weekdays and holidays each month. $\boldsymbol{z}_i$ is a vector of time-invariant variables including psychographics and demographics. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameter vectors. In the random effect model, μ_i and the explanatory variables are assumed to be independent.

4 Results

Our aim is to assess the impact of the existing consumer channel choice factors (RQ1) and contextual factors (RQ2) on online purchase amount using Tobit regression analysis. Table 4 shows the estimation result.

Concerning utilitarian benefits for acquisition utility, quality-consciousness appears to have a significant effect ($p < 0.01$), suggesting that quality-conscious consumers are more likely to purchase groceries online than those who are less quality-conscious. Further, the estimate of price-consciousness is also significant ($p < 0.01$). Therefore, we conclude that price-conscious consumers are more likely to purchase groceries online than those less price-conscious. In contrast, as hedonic benefits, innovativeness and shopping enjoyment appears to be insignificant ($p > 0.1$).

Regarding transaction utility, the estimate of time pressure is significant ($p < 0.01$). However, this result is contrary to the findings of previous studies [7, 24], which show that consumers with less time purchase online. This study indicates that consumers with lower time pressure are more likely to purchase groceries online than those with higher time pressure in the new normal environment. The estimate of search

Table 4 Parameter estimates

		coef	s.e	
	Constant	−12.207	1.658	**
Stay at home	Stay at home on weekdays	0.256	0.112	*
	Stay at home on holidays	0.113	0.175	
Psychographics	Quality consciousness	1.289	0.273	**
	Price consciousness	1.043	0.294	**
	Innovativeness	−0.101	0.278	
	Shopping enjoyment	0.288	0.266	
	Time pressure	−1.540	0.273	**
	Lower search cost online	1.115	0.255	**
	Store loyalty	−0.486	0.267	†
	Risk perceptions	−1.623	0.234	**
	Anxiety	0.590	0.292	*
	Impulsiveness	0.027	0.245	
Demographics	Gender (male = 1)	−5.276	0.564	**
	Age	0.120	0.022	**
	Number of family members	−1.221	0.274	**
	Income	0.313	0.097	**

**$p < 0.01$, *$p < 0.05$, †$p < 0.1$

cost online is positively significant ($p < 0.01$). Thus, consumers who perceive online search costs to be lower are more likely to purchase groceries online. Further, store loyalty is marginally significant ($p < 0.1$) with the expected sign. It was hypothesized that store loyalty would be strongly negative and significant, but the result indicates that store loyalty does not necessarily have a strong effect. The estimate of risk perception is negative and significant ($p < 0.01$). Therefore, consumers with lower perceived risk are more likely to purchase groceries online than those with higher perceived risk.

As the contextual factors after COVID-19 related to RQ2, we discover that anxiety is positive and significant ($p < 0.05$). Hence, consumers with high anxiety about COVID-19 are more likely to purchase groceries online. However, this anxiety-based purchasing is not as impulsive as panic buying. The estimate of impulsiveness is insignificant ($p > 0.1$).

Another perspective under the pandemic is how staying at home leads to online grocery shopping. The estimate of staying at home weekday is positive and significant ($p < 0.05$), but the estimate of holiday is insignificant ($p > 0.1$). This result indicates that consumers who feel the need to be home during the weekdays are more likely to do online grocery shopping.

5 Robustness Checks

We checked whether the insignificant explanatory variables could have created significant results for the other explanatory variables in the model [16]. After excluding all variables that showed insignificant results in Table 4, we estimated the model again. In the channel research context, [14] performs a robustness check in a similar way. As a result, even after excluding four variables—staying at home on holidays, innovativeness, shopping enjoyment and impulsiveness—the significance level and directions of the remaining explanatory variables did not change. Therefore, we confirmed the robustness of our results.

In addition, since this study used monthly data for six months, the time trend may have influenced the results. Therefore, we included dummy variables for the month in the model and re-estimated the model. As a result, none of the dummy variables for the month became significant, and the significance level and directions of the other explanatory variables did not change. Hence, we confirmed that our results are not influenced by any particular month.

6 Discussion

While the adoption of online shopping for groceries has remained lower than that in other retailing domains [24, 27], its popularity had been steadily increasing prior to the beginning of the COVID-19 pandemic. However, the long-lasting impact of COVID-19 has changed consumers' daily lifestyles and has accelerated online shopping. Under the new normal, better understanding the factors influencing consumers' online grocery shopping is helpful to various practitioners such as retailers, marketers and policymakers. Hence, this issue is influential for grocery retailing research [11]. We present an integrated approach to evaluate the impact of psychological factors on the consumers' actual purchasing behaviors. In previous studies, studies dealing with psychological factors and behavioral traits have been studied separately, but we propose a holistic view. Furthermore, this study captures the contextual factors of consumer behavior after COVID-19, which is the novelty of this study.

Among the existing consumer channel choice factors, we showed that utilitarian factors have more influence on online purchase amount than hedonic factors. One reason for this is that the activity of grocery shopping is considered a mundane, routine task in general [26]. Further, our result reveals that the convenience of online search is positively associated with online grocery shopping. The result reflects the consumers' tendency to make everyday purchases more efficiently.

Perceived risk is the influential factor in our results. While many studies have supported this [17, 26], recent research has suggested that this is not necessarily the case as online grocery shopping becomes more common [11]. However, the evidence shows that perceived risk is still strongly related to online grocery shopping in the

Japanese market. This tendency may depend on the maturity of the market in the target country.

An interesting finding is about time pressure. This study indicates that consumers with lower time pressure are more likely to purchase groceries online. However, previous studies have shown that consumers with less time are likely to purchase online for the purpose of time-saving [7, 11, 24, 26]. This study indicates that the opposite is true; rather, staying at home (especially on weekdays) and having more time to spare leads consumers to purchase more online. The self-restraint from going out due to COVID-19 could have given consumers more time to adopt a new shopping format. The dominance of physical stores in grocery shopping has made it difficult for online shopping to become widespread in the past, but with more time available, the new format may take root in the future.

Moreover, we find that consumers with higher anxiety about COVID-19 are more likely to purchase groceries online. Current and future retailing is expected to provide hygienic and efficient ways to shop for consumer convenience [13]. Therefore, this format of online grocery shopping is expected to become more popular, alleviating consumer anxiety.

In contrast, consumers' propensity to seek hedonically and entertainment benefits are not associated with online shopping. Further, the tendency for impulsive buying is not associated. Based on the results, the shift to online shopping does not encourage unplanned purchases. The critical question for retailers in the future would be whether online grocery retailing will increase the total amount of shopping by consumers, and if not, how to increase the total amount. According to [24], in the long run, multichannel grocery shoppers expand the share of wallets allocated to the online-visited chain compared to single-channel shoppers. However, there is a need for additional validation as to whether these phenomena are also occurring in the new normal environment. It will be a challenge for retailers to figure out how to make online shopping more enjoyable for consumers and increase unplanned purchases.

COVID-19 has forced people to change their lives, and retail formats have evolved along with it. Some of the changes could be permanent in the future. One of the reasons for the lack of online purchasing diffusion may be that the cost of switching channels was too high for consumers, either because of a lack of online experience or because of offline purchasing habits and inertia [20]. However, the pandemic forced consumers to try new formats. Moreover, as the effects of the pandemic lingered, it became a habit. Based on the empirical data, this study has implications regarding several important consumer characteristics. It is expected that many researchers and practitioners will use this knowledge to draw up strategies related to customer acquisition and retention after COVID-19.

7 Future Research

Our study has some limitations that point to areas of future research interest. The first limitation is the timeframe of our study, which includes only six months following the onset of COVID-19. A longitudinal study will be necessary in the future to identify whether consumers' online grocery shopping will become more popular. Second, our study focuses on online purchasing for all groceries using scanner panel data but does not consider any specific firm channels. Several studies of multichannel retailing provide implications regarding profitability and customer retention when a firm deploys multiple channels (e.g., [2, 7, 24]). It would be helpful to better understand how the factors identified in this study can be applied to the channel development of individual firms. Third, our online survey design has a limitation. We measured the stay home status variable for March 2020 through September 2020 in a single online survey. However, it is difficult for consumers to recall their past monthly stay home situation. To collect more accurate data, future studies should conduct monthly surveys.

Acknowledgements We appreciate INTAGE Inc. for permission to use the dataset.

References

1. Ailawadi, K.L., Neslin, S.A., Gedenk, K.: Pursuing the value-conscious consumer: Store brands versus national brand promotions. J. Mark. **65**(1), 71–89 (2001)
2. Ansari, A., Mela, C.F., Neslin, S.A.: Customer channel migration. J. Mark. Res. **45**(1), 60–76 (2008)
3. Babin, B.J., Darden, W.R., Griffin, M.: Work and/or fun: measuring hedonic and utilitarian shopping value. Journal of Consumer Research **20**(4), 644–657 (1994)
4. Bagozzi, R.P., Yi, Y.: On the evaluation of structural equation models. J. Acad. Mark. Sci. **16**(1), 74–94 (1988)
5. Briesch, R.A., Chintagunta, P.K., Fox, E.J.: How does assortment affect grocery store choice? J. Mark. Res. **46**, 176–189 (2009)
6. Campo, K., Breugelmans, E.: Buying groceries in brick and click stores: category allocation decisions and the moderating effect of online buying experience. J. Interact. Mark. **31**(3), 63–78 (2015)
7. Chintagunta, P.K., Chu, J., Cebollada, J.: Quantifying transaction costs in online/off-line grocery channel choice. Mark. Sci. **31**(1), 96–114 (2012)
8. Chiu, H.C., Hsieh, Y.C., Roan, J., Tseng, K.J., Hsieh, J.K.: The challenge for multichannel services: cross-channel free-riding behavior. Electron. Commer. Res. Appl. **10**(2), 268–277 (2011)
9. Chu, J., Arce-Urriza, M., Cebollada-Calvo, J.J., Chintagunta, P.K.: An empirical analysis of shopping behavior across online and offline channels for grocery products: the moderating effects of household and product characteristics. J. Interact. Mark. **24**(4), 251–268 (2010)
10. Fornell, C., Larcker, D.F.: Structural equation models with unobservable variables and measurement error: algebra and statistics. J. Mark. Res. **18**(3), 328–388 (1981)
11. Frank, D.A., Peschel, A.O.: Sweetening the deal: the ingredients that drive consumer adoption of online grocery shopping. J. Food Prod. Mark. **26**(8), 535–544 (2020)

12. Gao, X., Shi, X., Guo, H., Liu, Y.: To buy or not buy food online: the impact of the COVID-19 epidemic on the adoption of e-commerce in China. PLoS ONE **15**(8), e0237900 (2020)
13. Gauri, D.K., Jindal, R.P., Ratchford, B., Fox, E., Bhatnagar, A., Pandey, A., Navallo, J.R., Fogarty, C.S., Howerton, E.: Evolution of retail formats: past, present, and future. J. Retail. **97**(1), 42–61 (2021)
14. Gensler, S., Neslin, S.A., Verhoef, P.C.: The showrooming phenomenon: it's more than just about price. J. Interact. Mark. **38**, 29–43 (2017)
15. Grashuis, J., Skevas, T., Segovia, M.S.: Grocery shopping preferences during the COVID-19 pandemic. Sustainability **12**(13), 5369 (2020)
16. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis, 7th edn. Prentice Hall (2010)
17. Hansen, T.: Determinants of consumers' repeat online buying of groceries. Int. Rev. Retail Distrib. Consum. Res. **16**(1), 93–114 (2006)
18. Jensen, K.L., Yenerall, J., Chen, X., Yu, T.E.: US consumers' online shopping behaviors and intentions during and after the COVID-19 pandemic. J. Agric. Appl. Econ. **53**, 416–434 (2021)
19. Jepsen, A.L.: Factors affecting consumer use of the internet for information search. J. Interact. Mark. **21**(3), 21–34 (2007)
20. Kannan, P.K., Kulkarni, G.: The impact of Covid-19 on customer journeys: implications for interactive marketing. J. Res. Interact. Mark. **16**(1), 22–36 (2022)
21. Konuş, U., Verhoef, P.C., Neslin, S.A.: Multichannel shopper segments and their covariates. J. Retail. **84**(4), 398–413 (2008)
22. Li, J., Konuş, U., Pauwels, K., Langerak, F.: The hare and the tortoise: do earlier adopters of online channels purchase more? J. Retail. **91**(2), 289–308 (2015)
23. Melis, K., Campo, K., Breugelmans, E., Lamey, L.: The impact of the multi-channel retail mix on online store choice: does online experience matter? J. Retail. **91**(2), 272–288 (2015)
24. Melis, K., Campo, K., Lamey, L., Breugelmans, E.: A bigger slice of the multichannel grocery pie: when does consumers' online channel use expand retailers' share of wallet? J. Retail. **92**(3), 268–286 (2016)
25. Ministry of Economy, Trade and Industry: FY2020 e-commerce market survey. https://www.meti.go.jp/english/press/2021/0730_002.html. Accessed 3 Mar 2022
26. Mortimer, G., Fazal, H.S., Andrews, L., Martin, J.: Online grocery shopping: the impact of shopping frequency on perceived risk. Int. Rev. Retail Distrib. Consum. Res. **26**(2), 202–223 (2016)
27. Nakano, S., Kondo, F.N.: Customer segmentation with purchase channels and media touch-points using single source panel data. J. Retail. Consum. Serv. **41**, 142–152 (2018)
28. Nakano, S., Akamatsu, N., Mizuno, M.: Consumer panic buying: understanding the behavioral and psychological aspects. SSRN 3796825 (2021)
29. Neslin, S.A.: The omnichannel continuum: integrating online and offline channels along the customer journey. J. Retail. **98**(1), 111–132 (2022)
30. Neslin, S.A., Shankar, V.: Key issues in multichannel customer management: current knowledge and future directions. J. Interact. Mark. **23**(1), 70–81 (2009)
31. Roggeveen, A.L., Sethuraman, R.: How the COVID-19 pandemic may change the world of retailing. J. Retail. **96**(2), 169–171 (2020)
32. Sheth, J.: Impact of Covid-19 on consumer behavior: will the old habits return or die? J. Bus. Res. **117**, 280–283 (2020)
33. Verhoef, P.C., Neslin, S.A., Vroomen, B.: Multichannel customer management: understanding the research-shopper phenomenon. Int. J. Res. Mark. **24**(2), 129–148 (2007)
34. Vroegrijk, M., Gijsbrechts, E., Campo, K.: Close encounter with the hard discounter entry: a multiple-store shopping perspective on the impact of local hard-discounter entry. J. Mark. Res. **50**(5), 606–626 (2013)
35. Winterich, K.P., Haws, K.L.: Helpful hopefulness: the effect of future positive emotions on consumption. J. Consum. Res. **38**(3), 505–524 (2011)

Skeletal Muscle Segmentation at the Third Lumbar Vertebral Level in Radiotherapy CT Images

Xuzhi Zhao, Haizhen Yue, Yi Du, Shuang Hou, Weiwei Du, and Yahui Peng

Abstract A computer algorithm is proposed to segment skeletal muscles at the third lumbar vertebral (L3) level in radiotherapy computed tomography (CT) images. Included in the study are 20 patients who were diagnosed with rectal cancer and their pelvic CT images were acquired with a radiotherapy CT scanner. An oncologist selects an axial CT slice at the L3 level for each patient and manually annotates the ground truth, or the skeletal muscles, including both the abdominal and the paraspinal muscles. The abdominal muscles are segmented with morphological techniques while the paraspinal muscles are segmented with an approach including an adaptive thresholding method, connected component analysis, and morphological techniques. Both Dice similarity coefficient and 95th percentile of the Hausdorff distance are used to assess the segmentation accuracy, and they are 93.8 $\pm$ 1.6 % and 4.8 $\pm$ 1.2 mm, respectively, averaged over the entire dataset. In conclusion, the proposed algorithm is demonstrated to be accurate in segmenting skeletal muscles at the L3 level in radiotherapy CT images.

X. Zhao · S. Hou · Y. Peng (✉)
School of Electronic and information Engineering, Beijing Jiaotong University, Beijng, China
e-mail: yhpeng@bjtu.edu.cn

X. Zhao
e-mail: 20111055@bjtu.edu.cn

S. Hou
e-mail: 21120007@bjtu.edu.cn

H. Yue · Y. Du
Department of Radiation Oncology, Peking University Cancer Hospital and Institute, Beijng, China
e-mail: haizhenyue@bjcancer.org

Y. Du
e-mail: yidu@bjcancer.org

W. Du
Department of Information Science, Kyoto Institute of Technology, Kyoto, Japan
e-mail: duweiwei@kit.ac.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_6

Keywords Skeletal muscles · Muscle segmentation · Third lumbar vertebral (L3) level · CT images · Radiotherapy

1 Introduction

Rectal cancer is one of the common malignant neoplasm diseases in the world [1] and the nutritional status of patients is of great significance to the prognosis outcome [2]. Skeletal muscle mass at the third lumbar vertebral (L3) level has been shown to correlate with patient nutritional status [3, 4]. Computed tomography (CT) plays an import role in muscle mass evaluation since the morphology of the muscles can be demonstrated clearly [5]. How to develop a computer program to segment skeletal muscles accurately at the L3 level in CT images has been investigated extensively [6–13].

Previous studies on the segmentation of skeletal muscles at the L3 level in CT images revolved around traditional image processing algorithms [6–9] and deep learning-based methods [10–13]. Traditional algorithms include shape prior modeling-based image registration [6, 9], atlas-based segmentation [7] and fuzzy c-means clustering [8]. Deep learning-based methods include fully convolutional neural networks [10, 11], UNet-based models [10, 12] and YOLOv3-based models [13]. However, the existing traditional algorithms tend to rely on complex modeling and thus, having trouble reproducing the results on different datasets, while deep learning-based methods generally require a large number of images with labels manually annotated by skilled radiologists for network training.

In this study, a novel computer algorithm is proposed to segment skeletal muscles at the L3 level in radiotherapy CT images. The algorithm is based on the morphological characteristics of skeletal muscles in radiotherapy CT images. Because the algorithm only consists of basic image processing techniques, it does not require a large number of labeled images for training.

2 Materials and Methods

2.1 Patients

Included in the study are 20 rectal cancer patients (15 males and 5 females) who were admitted to the Department of Radiation Oncology at Peking University Cancer Hospital & Institute from April 2015 to December 2016. The patient age ranges from 31 to 77 years, with the mean of 61 years.

Table 1 CT imaging parameters

Imaging parameter	Value
Tube voltage (kVp)	120
Tube current (mAs)	190
Slice thickness (mm)	5
Matrix size	512X512
Pixel spacing (mm)	(1.27, 1.27)

2.2 CT Image Acquisition

All patients underwent both plain and contrast-enhanced pelvic CT studies with a Sensation Open CT scanner (Siemens Healthineers, Erlangen, Germany) for radiotherapy. The corresponding CT imaging parameters are shown in Table 1.

Only the plain pelvic CT images are used for muscle segmentation in this study.

2.3 Ground Truth

On the axial CT images at the L3 level, an oncologist selects a best slice that demonstrates the structure of the spine morphology for each patient and annotates the skeletal muscles manually with the ITK-SNAP software [14]. A total of 20 axial images are annotated and used as the ground truth for the assessment of the segmentation accuracy.

Figure 1a shows a selected radiotherapy CT image at the L3 level and Fig. 1b shows the ground truth annotated by the oncologist.

2.4 Skeletal Muscle Segmentation

The flowchart of the proposed skeletal muscle segmentation algorithm is shown in Fig. 2. The abdominal and paraspinal muscles are segmented separately after preprocessing [15].

2.4.1 Preprocessing

- *Global thresholding*: The input image is first segmented with a given pair of lower and upper thresholds, −29 and 150 Hounsfield unit (HU), respectively, which are considered the range of standard skeletal muscle CT values [16]. This step removes most pixels belonging to subcutaneous and visceral adipose tissue.

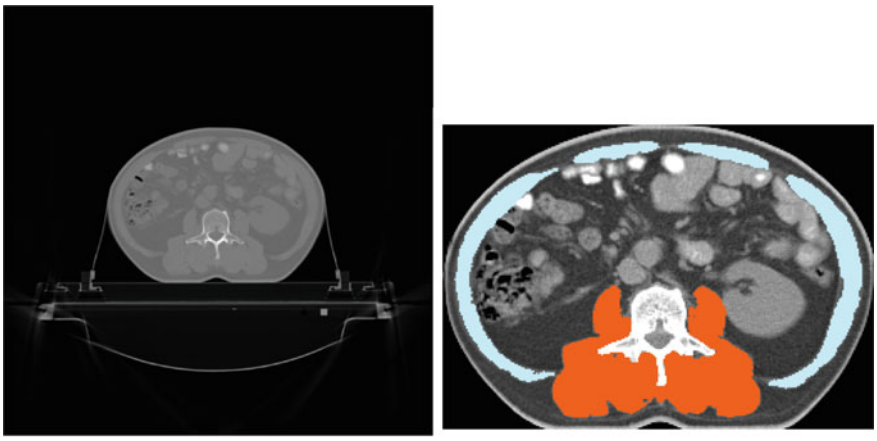


Fig. 1 A radiotherapy CT image at the level of the third lumbar vertebral (**a**) and the corresponding manually-annotated skeletal muscles (**b**), including both the abdominal muscles (blue) and the paraspinal muscles (orange), given by an oncologist. Note that, for better visualization, the annotated image is cropped, and the gray level is adjusted

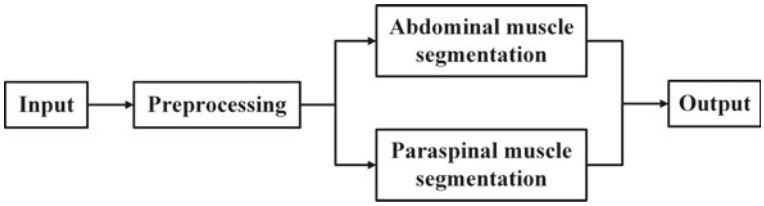


Fig. 2 Flowchart of the proposed algorithm

- *Skin removal*: The outermost pixels of skin tissue is removed through connected component analysis and morphology processing.

2.4.2 Abdominal Muscle Segmentation

- *Convex hull extraction*: The outer contour of the segmented region is found, then the convex hull of the outer contour is extracted (Fig. 3).
- *Distance map generation*: To identify the thin layer of the abdominal muscles, a Chebyshev distance map is generated from the convex hull (Fig. 4).
- *Abdominal muscle identification*: Inside the segmented region, the most probable distance of the segmented pixels from the convex hull is recorded (Fig. 5). Segmented pixels whose corresponding distance greater than or equal to the most probable distance are removed.

Fig. 3 An example showing the outer contour (red) and the corresponding convex hull (green) of the muscles

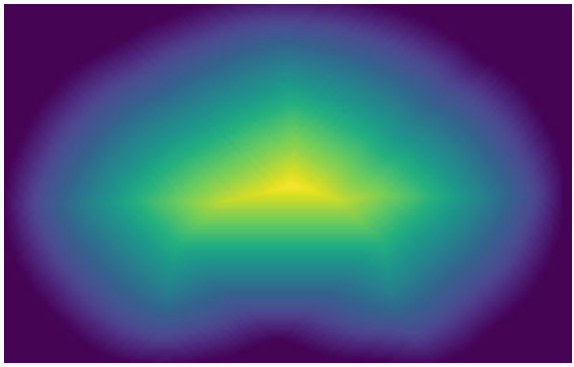


Fig. 4 The Chebyshev distance map inside the convex hull

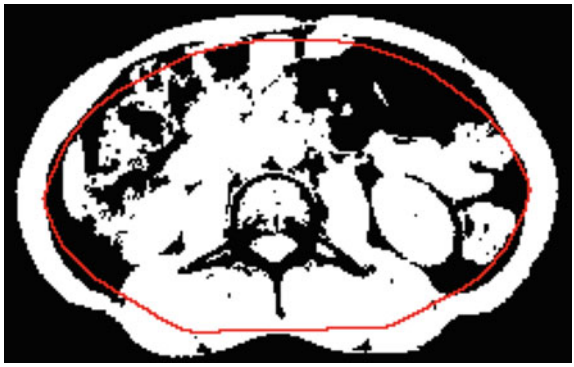


Fig. 5 Using the most probable distance to estimate the inside boundary of the abdominal muscles (red contour)

- *Abdominal muscle refinement*: To refine the inner side of the region obtained in the previous step, the convex hull of the inner contour of the previous result is extracted (Fig. 6). Segmented pixels inside the convex hull are removed. The updated inner side is repeatedly refined using the same method until the number of removed pixels is less than 150.

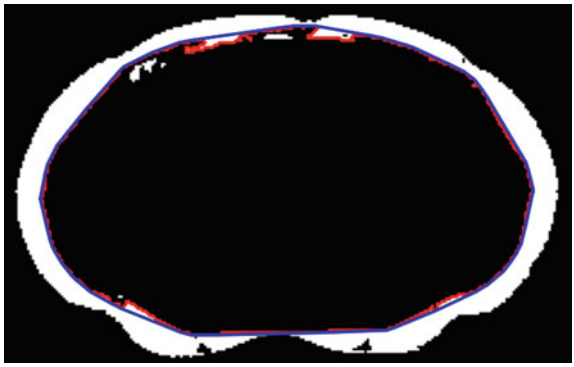


Fig. 6 The convex hull (blue) of the original inner contour (red) refines the inner profile of the abdominal muscles

2.4.3 Paraspinal Muscle Segmentation

- *Adaptive thresholding*: Caring about pixels of the input image that lie within the region of the preprocessed result, the histogram profile is fitted to a normal distribution to obtain the mean value μ and the standard deviation σ. Adaptive thresholds are then set to be $\mu - 1.5\sigma$ and $\mu + 1.5\sigma$ (Fig. 7), which are used as the lower and upper limits. Pixels whose value in between are segmented.
- *Paraspinal muscle localization*: The L3 vertebral body is segmented using a threshold of 110HU. Then, the bounding box of the vertebral body is identified. An enlarged region is cropped after dilating the upper bound by three pixels and extending the left and right bounds to evenly double the width (Fig. 8).
- *Noise removal*: To remove noise in the previous result, connected components analysis is used to remove small regions. The largest eight connected components are retained.
- *Paraspinal muscle identification*: A series of rectangular boxes are generated adaptively in the upper left and upper right regions of the cropped image in the following way. Some fixed-length lines in vertical direction are set in the upper left (right) region of the image and the end part of each line is checked whether intersects with muscle tissue. If it intersects, skip. If not, a horizontal line is generated from the left (right) border of the image to the end of the vertical line. Then, the spatial position of the horizontal line lowers continuously along the y-axis until its end part intersects with muscle tissue. By removing the pixels inside the rectangular regions, the paraspinal muscles are roughly identified (Fig. 9).
- *Paraspinal muscle refinement*: The connected component analysis is then used. Criteria regarding the size and location of the regions are enforced to filter out the connected components of non-muscle tissue. Regions inside the vertebral body are removed. Regions with size smaller than or equal to 8 are removed. Small regions near the top, left, and right border of the image are removed. The holes in the image are filled.

Finally, the results of the abdominal muscle segmentation and the paraspinal muscle segmentation are combined to get the complete skeletal muscles.

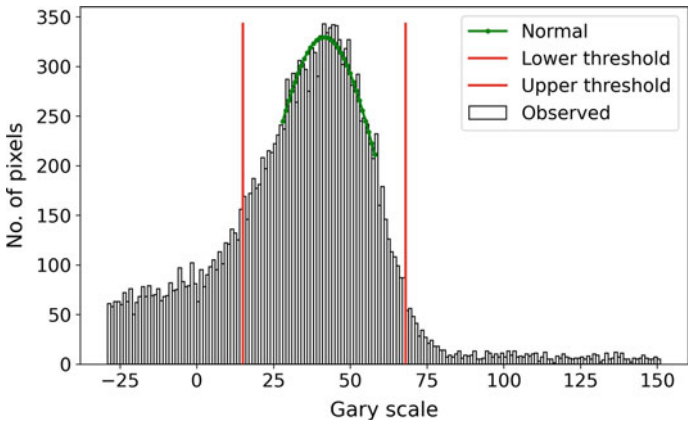


Fig. 7 The histogram in a certain range is fitted to a normal distribution (green) to determine the adaptive lower and upper thresholds (red) for paraspinal muscle segmentation

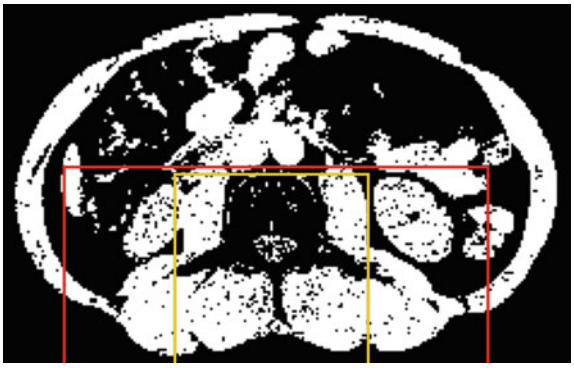


Fig. 8 Using the bounding box of the vertebral body (yellow) to localize the region of paraspinal muscles (red)

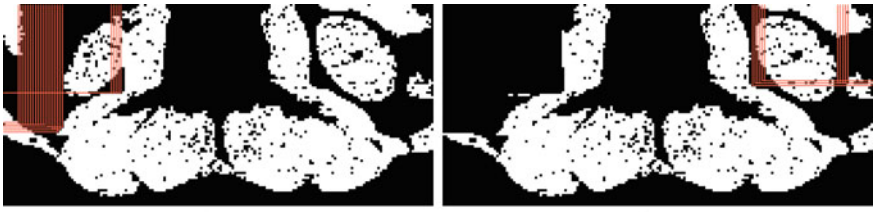


Fig. 9 By removing non-muscle pixels on the left **a** and right **b** upper regions, paraspinal muscles are identified

Fig. 10 The abdominal muscle segmentation result

Fig. 11 The paraspinal muscle segmentation result

2.5 Evaluation Metrics

To assess the segmentation accuracy, Dice similarity coefficient (DSC) and 95th percentile of the Hausdorff distance (HD95) are used [17]. The DSC is a measure of pixel-wise overlap of the segmented and reference regions and HD95 evaluates the distance between segmented and reference boundaries.

3 Results

The segmented abdominal muscles, paraspinal muscles and complete skeletal muscles are shown in Figs. 10, 11, and 12, respectively.

Two segmentation results of skeletal muscles at the L3 level are shown in Fig. 13, representing the best and worst segmentation performance in terms of DSC. HD95 is also given for reference.

The histograms of segmentation accuracy are compared between results from using simple global thresholding method and the proposed algorithm (Fig. 14). The statistics of the histograms are shown in Table 2. It is clear that the proposed algorithm improves the segmentation performance on top of the global thresholding method.

Fig. 12 The complete skeletal muscle segmentation result

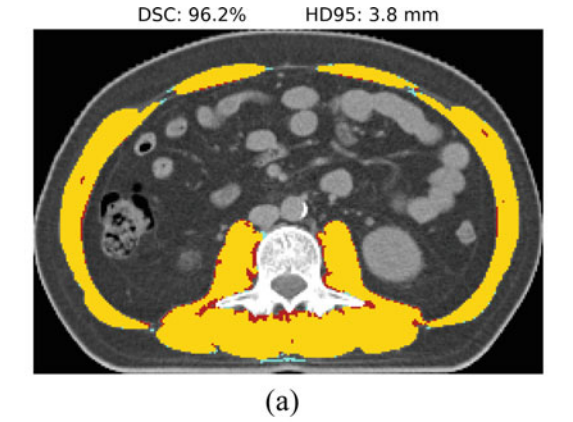


Fig. 13 Demonstration of the best (**a**) and worst (**b**) skeletal muscle segmentation results from using the proposed algorithm. The proposed skeletal muscle segmentation result (cyan) and manual segmentation result (red) are highly overlapped (yellow). DSC: Dice similarity coefficient. HD95: 95th percentile of the Hausdorff distance

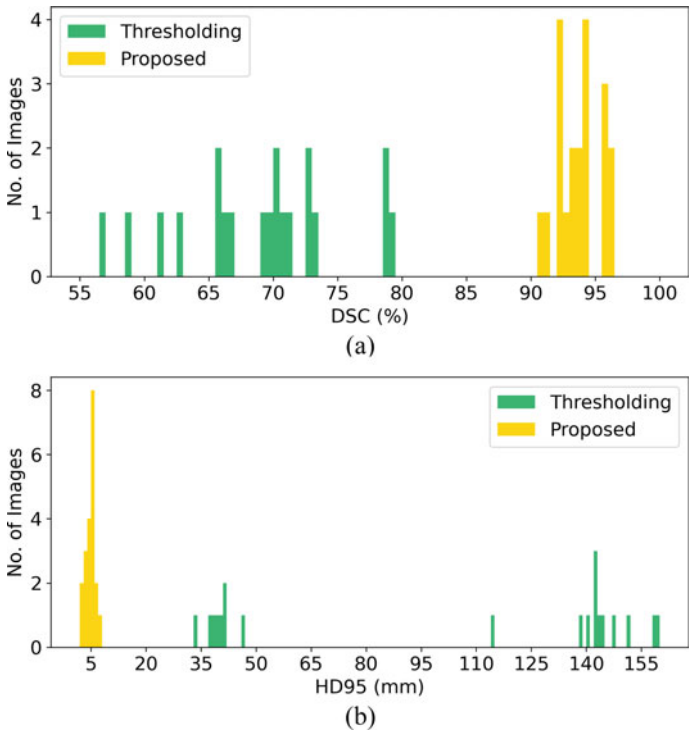


Fig. 14 Histograms of Dice similarity coefficient (**a**) and 95th percentile of the Hausdorff distance (**b**) over the entire dataset for the global thresholding method and the proposed segmentation algorithm

Table 2 Comparison of segmentation results obtained using different methods. All values are reported as MEAN ± SD

Method	Evaluation Metric	
	DSC (%)	HD95 (mm)
Thresholding	69.1 ± 6.2	102.2 ± 52.9
Proposed	93.8 ± 1.6	4.8 ± 1.2

4 Discussion

The nutrition status of rectal cancer patients is closely related to the prognosis of the disease [2]. Features derived from skeletal muscles at the L3 level are correlated to the nutritional status of the patients [3, 4]. Hence, it is of great importance to segment skeletal muscles accurately at the L3 level in CT images.

The average DSC of the proposed algorithm is 93.8 ± 1.6 on the dataset in study. For comparison, previous studies reported DCS of 94.53 ± 5.06 [9], 98.11 ± 1.47

[10], and 98 [13], not on the same dataset. Although the mean segmentation accuracy of the proposed algorithm is slightly lower than similar studies [9, 10, 13], the proposed algorithm is much less expensive. The proposed algorithm circumvents the drawbacks of complex modeling and the need for a large number of images with annotations manually labeled for network training. In addition, the proposed method worked in radiotherapy CT images, whose quality is normally not as good as diagnostic CT images.

There are limitations of this study. First, the number of cases is relatively small. The segmentation accuracy of the proposed algorithm is estimated using 20 radiotherapy CT images. More cases can be used to improve the reliability of the estimation. Second, the proposed algorithm is tested on plain CT images acquired with a single radiotherapy device. In future studies, the proposed algorithm will be tested on more imaging devices and other imaging modes.

5 Conclusion

In conclusion, the promising results indicate that the proposed image-processing algorithm is accurate in segmenting skeletal muscles at the L3 level in radiotherapy CT images. It might be served as an annotation tool for the assessment of nutritional status of rectal cancer patients.

Acknowledgements This study was partially supported by the Beijing Natural Science Foundation (No. 1212011, 1202009) and National Natural Science Foundation of China (No. 12005007).

References

1. Valentini, V., Beets-Tan, R., Borras, J.M., Krivokapi, Z., Verfaillie, C.: Evidence and research in rectal cancer. Radiother. Oncol. **87**(3), 449–474 (2008)
2. Aleksandrova, K., et al.: Metabolic syndrome and risks of colon and rectal cancer: The European prospective investigation into cancer and nutrition study. Cancer Prev. Res. **4**(11), 1873–1883 (2011)
3. Wang, S., et al.: The value of L3 skeletal muscle index in evaluating preoperative nutritional risk and long-term prognosis in colorectal cancer patients. Sci. Rep. **10**(1), 1–11 (2020)
4. Bamba, S., et al.: Assessment of body composition from CT images at the level of the third lumbar vertebra in inflammatory bowel disease. Inflamm. Bowel Dis. **27**(9), 1435–1442 (2021)
5. Goldman, L.W.: Principles of CT and CT technology. J. Nucl. Med. Technol. **35**(3), 115–128 (2007)
6. Chung, H., Cobzas, D., Birdsell, L., Lieffers, J., Baracos, V.: Automated segmentation of muscle and adipose tissue on CT images for human body composition analysis. In: Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling, vol. 7261, pp. 197–204 (2009)
7. Meesters, S. et al.: Multi atlas-based muscle segmentation in abdominal CT images with varying field of view. In: International Forum on Medical Imaging in Asia (IFMIA), pp. 16–17 (2012)
8. Wei, Y., Tao, X., Xu, B., Castelein, A.: Paraspinal muscle segmentation in CT images using GSM-based fuzzy C-means clustering. J. Comput. Commun. **2**(9), 70–77 (2014)

9. Popuri, K., Cobzas, D., Esfandiari, N., Baracos, V., Jägersand, M.: Body composition assessment in axial CT images using FEM-based automatic segmentation of skeletal muscle. IEEE Trans. Med. Imaging **35**(2), 512–520 (2015)
10. Dabiri, S., Popuri, K., Feliciano, E.M.C., Caan, B.J., Baracos, V.E., Beg, M.F.: Muscle segmentation in axial computed tomography (CT) images at the lumbar (L3) and thoracic (T4) levels for body composition analysis. Comput. Med. Imaging Graph. **75**, 47–55 (2019)
11. Liu, Y., Zhou, J., Chen, S., Liu, L.: Muscle segmentation of L3 slice in abdomen CT images based on fully convolutional networks. In: 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–5 (2019)
12. Dabiri, S., et al.: Deep learning method for localization and segmentation of abdominal CT. Comput. Med. Imaging Graph. **85**, 101776 (2020)
13. Ha, J., et al.: Development of a fully automatic deep learning system for L3 selection and body composition assessment on computed tomography. Sci. Rep. **11**(1), 1–12 (2021)
14. Yushkevich, P.A., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006)
15. Engelke, K., Museyko, O., Wang, L., Laredo, J.-D.: Quantitative analysis of skeletal muscle by computed tomography imaging-state of the art. J. Orthop. Transl. **15**, 91–103 (2018)
16. Amini, B., Boyle, S.P., Boutin, R.D., Lenchik, L.: Approaches to assessment of muscle mass and myosteatosis on computed tomography: A systematic review. J. Gerontol. A Biol. Sci. Med. Sci. **74**(10), 1671–1678 (2019)
17. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)

Speed-Up Single Shot Detector on GPU with CUDA

Chenyu Wang, Toshio Endo, Takahiro Hirofuchi, and Tsutomu Ikegami

Abstract Nowadays, most of the current research on object detection is to improve the whole framework, in order to improve the accuracy of detection, but another problem of object detection is the detection speed. The more complex the architecture, the slower the speed. This time, we implemented a Single Shot Multibox Detector(SSD) using GPU with CUDA. We have improved the object detection speed of SSD, which is one of the most regularly used objects detection frameworks. The most time-consuming part, the VGG16 network, is rewritten by using cuDNN, which is made faster by about 9%. The second time-consuming part is post-processing, where non-maximum-suppression (NMS) is performed. We accelerated NMS by implementing our new algorithms that are suitable for GPUs, which is about 52% faster than the original PyTorch version [16]. We also ported those parts that were originally executed on CPU to GPU. In total, our GPU-accelerated SSD can detect objects 22.5% faster than the original version. We demonstrate that using GPUs to accelerate existing frameworks is a viable approach.

Keywords Object detection · GPU · CUDA · Single shot multibox detector · Parallel algorithm

C. Wang (✉)
Tokyo Institute of Technology & National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
e-mail: wang.c.ao@m.titech.ac.jp

T. Endo
Tokyo Institute of Technology, Tokyo, Japan
e-mail: endo@is.titech.ac.jp

T. Hirofuchi · T. Ikegami
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
e-mail: t.hirofuchi@aist.go.jp

T. Ikegami
e-mail: t-ikegami@aist.go.jp

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_7

1 Introduction

1.1 Context

Object detection is a type of computer vision that recognizes real-world objects in an image or video sequence. Humans have an innate ability to recognize objects: despite the fact that an object might seem different depending on the viewpoint, scale, translation, or rotation, a human can identify it with minimal effort. Even if an item is partially obscured, it may sometimes be recognized. Our ultimate objective is to discover objections to a contemporary computing system with human precision as well as computer speed. In the realm of computer vision, this problem is still difficult. Artificial Neural Networks (ANNs), which are inspired by our biological nervous system, have been presented as a way to educate computers to learn in the same way that people do. Artificial neural networks are constructed by individual "neurons" with weights that mimic the connectivity of neurons, similar to biological neural networks. Training is the process of determining each neuron's weight, while inference is the process of utilizing a trained neural network to do a computing job. An ANN is made up of at least two layers, input, and output, each of which may contain many perceptrons. The "hidden layer" is the layer that sits between the input and output layers. An ANN with several hidden layers is known as a deep neural network.

With the rapid development of object detection, the commercial value of object detection has been found in many areas, such as massive security surveillance and autonomous driving. Many commercial companies have launched their products or services based on object detection and the market keeps growing. However, an efficient and reliable object detection system is yet to be found on the market. Because for commercial use, the speed of detection has always been an issue [13]. Although object detection using deep convolutional neural networks has advanced dramatically in recent years, there is still a long way to go to achieve commercial use to ensure the accuracy rate, we must also achieve real-time detection. Object detection is the process of identifying whether or not an object of interest is present in a picture or video frame, independent of its size, orientation, or surroundings. There are numerous techniques for performing detection operations, the simplest of which is the template matching approach employing a moving window slide, like R-CNN [9]. However, the method's calculating time is a significant issue.

Due to the rising resolutions of both image and video, sequential processing of high-resolution pictures and videos using a single core processor cannot satisfy the needed speedup to detect an item. As a result, modern academics are focusing on parallel processing, which involves doing several calculations at the same time utilizing either a software paradigm such as a GPU or a soft-hardware paradigm such as FPGA. Both of the paradigms have greater performance than a single core CPU. The GPU may be utilized as a general-purpose accelerator. The Compute Unified Device Architecture (CUDA) technology included with NVIDIA products is now frequently used to speed up image/video processing applications rather than graph-

ics. In 2010, Mehta et al. presented a novel [20], high-performance implementation of SAD(The Sum of Absolute Differences), which is a measure of the similarity between image blocks, on the general purpose GPU architecture using NVIDIA's CUDA [1]. Their research proved that object detection can achieve faster detection speed through GPU acceleration, and can have a good effect. Real-time execution of very complicated computer vision algorithms has become a reality because of recent breakthroughs in GPU computing speed. Multithreaded data-parallel GPU architectures are commonly used in applications such as advanced driver assistance systems (ADAS), autonomous driving, scene interpretation, intelligent video analytics, and facial recognition, among others. It also has a lot of potential in the future, thus it's a terrific study topic.

1.2 Motivation

One of the most often utilized detection networks is SSD(Single Shot Multibox Detector) [19]. When compared to other high-speed detection networks, it achieves object detection with a high mean Average Precision(mAP). At the same time, SSD has achieved a very good balance between accuracy and detection speed. It seems to be an excellent research object. Major object identification networks are built on convolutional neural networks, and numerous approaches have been developed to speed up the computation of CNN's convolution and full connection layers. However, in detection networks, not only convolution and complete connection but also other modules, need a significant amount of processing time. Therefore, in addition to the CNN layer, attention should also be paid to the calculation of other layers. If we accelerate other layers by GPU, we can also achieve a faster detection speed, which is a major goal of our research.

1.3 Our Contribution

Our main contribution is SSD512 detection acceleration through CUDA, which is about 22.5% faster than the normal version [16]. We implemented all the implementation using CUDA [25] and C++ and rewrite Backbone (VGG16) with cuDNN, which made backbone about 9% faster than the original version. Based on the characteristics of GPU, we processed the image and classification in parallel. Finally, we proposed a CUDA-based Non-maximum Suppression(NMS) algorithm, which is about 52% faster than the original version.

2 Related Work

In this chapter, I will introduce the contents related to our research.

2.1 Object Detection

Since 2010, manual feature extraction-based object identification algorithms have been stagnant. Region selection, feature extraction, and classification are the three steps in the classical technique. Furthermore, in object detection contests such as the "ImageNet Large Scale Visual Recognition Challenge (ILSVRC)" [24], "COCO: Common Objects in Context Detection Challenge" [18] and "PASCAL VOC: The PASCAL Visual Object Classes Challenge" [6], approaches using deep neural networks received the greatest scores. However, this pipeline has two problems: a good region-selection method with a simple time complexity has yet to be discovered, and the robustness of manual feature extraction cannot be guaranteed. A new method [15] based on convolutional neural networks is introduced in 2012, which started a new era to object detection. CNN significantly improved both issues, particularly for approaches based on region suggestion called R-CNN [9]. Modern approaches are divided into two categories: one-stage frameworks and two-stage frameworks. YOLO [22], SSD [19], and RetinaNet [17] are examples of one-stage frameworks. Faster R-CNN[14], Mask R-CNN [10], and Cascade R-CNN [3] are examples of two-stage frameworks.

- **Two-stage frameworks**: The two-stage frameworks has higher accuracy, but the detection speed will be relatively slower.
- **One-stage frameworks**: The one-stage frameworks has faster detection speed, but the accuracy will be relatively lower.

Why did this happen? The reason for this is that the ratio of background to non-background areas are not balanced. In the two-stage frameworks, background samples are screened through the proposal stage, and the number of candidates is firstly reduced. There is no one statistic that can determine if two-stage frameworks are superior to one-stage frameworks, or vice versa. With varying degrees of success, both branches attempt to strike a balance between speed and precision. Two-stage frameworks represent the pinnacle of object detection accuracy, but one-stage frameworks are better suited to real-time circumstances such as video surveillance or high-speed motion capture. In the later classification phase, the ratio of the number of background and off-background fields is fixed at 3:1, with Online Hard Example Mining(OHEM) [26] to maintain balance. In the one-stage frameworks, there are too many fields simply classified as background, and these methods are very inefficient. Therefore, RetinaNet proposed to introduce a new Loss function, Focal Loss, to solve this problem.

A region proposal network (RPN) is used in two-stage frameworks to offer possible bounding boxes, which the detecting head network uses to refine the bounding box

coordinates and forecast classification scores. In a single network pass, the one-stage frameworks predict the box coordinates and classification result. Non-maximum suppression (NMS) is one of the most important post-processing procedures for removing boxes with comparable locations and shapes but with lower confidence levels. In [13], they analyzed the comparison of various object detection frameworks. They not only analyzed the correctness of different framework detection but also specifically analyzed the detection speed of various frameworks, which is also an important indicator of the maturity of the framework.

2.2 CUDA

CUDA [1] is a parallel computing platform and application programming interface (API) that enables applications to use specific types of graphics processing units (GPUs) for general-purpose processing (a technique known as general-purpose computing on GPUs) (GPGPU). CUDA is a software layer that allows computing kernels to have direct access to the GPU's virtual instruction set and parallel computational units. C, C++, and Fortran are among the programming languages supported by CUDA. In contrast to the previous APIs like Direct3D and OpenGL, which needed significant graphics programming abilities, this accessibility makes it easy for parallel programming experts to exploit GPU resources [25]. By compiling such code to CUDA, CUDA-powered GPUs also support programming frameworks like OpenMP, OpenACC, and OpenCL. There are several advantages that give CUDA an edge over traditional general-purpose graphics processor (GPGPU) computers with graphics APIs:

- Unified memory and unified virtual memory
- Shared memory:provides a faster area of shared memory for CUDA threads
- Scattered reads: code can be read from any address in memory
- Improved performance to transport data between CPU and GPU
- There is full support for bitwise and integer operations

2.3 Single Shot Multibox Detector

Single Shot Multibox Detector (SSD) is a well-known one-stage framework. In R-CNN, the bounding box is moved in various ways and the calculation is performed by CNN each time, so it took a considerable amount of processing time to detect an object from a single image. On the other hand, SSD performs both "regional candidate detection" and "classification" of objects in a single CNN operation, as the name "Single Shot" implies. This has made it possible to speed up the object detection process. SSD is designed to output multi-scale boxes from various output layers. The model architecture is shown below in Fig. 1. According to the size of

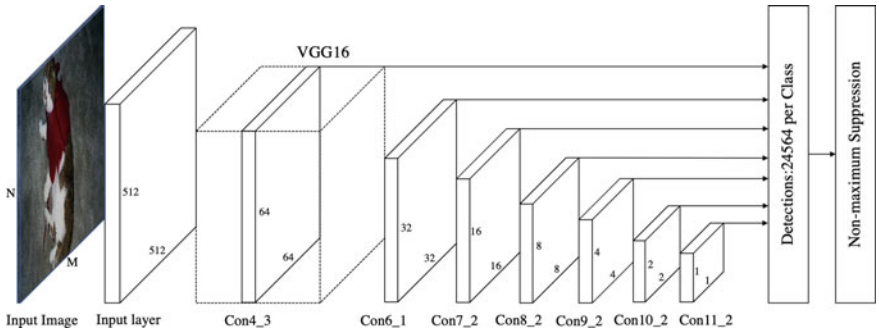


Fig. 1 Single Shot Multibox Detector Model(SSD512)

the input image, the Single Shot Multibox detector is mainly divided into two types, SSD300 [19] and SSD512, shown in Fig. 1. The basic architecture of SSD, as shown in Fig. 1, is based on VGG16 [27]. The layers of the first part of the network are constructed according to the standard architecture for image classification, which is called the base network. Eigenvalues are detected in the underlying network. The subsequent layers are auxiliary structures for multiple object detection through multi-scale feature maps. A total of 24,564 candidate boxes are generated in SSD512, which is compared with 8,732 in SSD300. Because of the increased size of an input image, a larger size of feature map is generated, and more candidate regions are generated. This number of candidate boxes are sieved by NMS algorithm.

2.4 Non-maximum Suppression

Non Maximum Suppression (NMS) is a technique used in numerous computer vision tasks. It is a class of algorithms to select one entity such as bounding boxes, out of many overlapping entities. We can choose the selection criteria to arrive at the desired results. The criteria are most commonly some form of probability number and some form of overlap measure such as Intersection over Union. It is especially true for high-resolution image detection, where more candidate regions are generated and the calculation time of NMS increases. Among all object detection frameworks, Non-maximum Suppression(NMS) is also an extremely important algorithm, which greatly affects the accuracy and speed of detection. Especially at the present, for high-resolution image detection, not only higher accuracy is achieved, but also more bounding boxes are generated, so the calculation time of NMS gradually increases.

To improve detection accuracy, different NMS variants have been proposed [2, 11, 12]. Instead of discarding boxes, Soft NMS [2] reduces the confidence score as a continuous function of intersection-union (IoU) and keeps all boxes. Continuous functions are hand-designed in [2], and a special network is learned in [12] to regain confidence. In addition to changing the confidence, in [11], the bounding box coor-

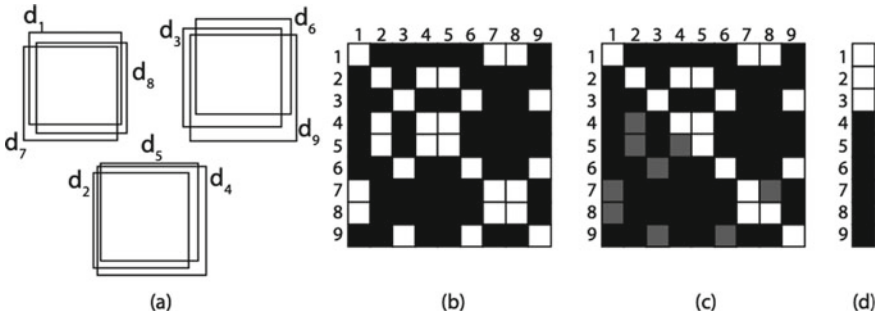


Fig. 2 Visualization of GPU-NMS proposal. **a** example candidates generated by a detector (3 objects, 9 detections); boolean matrix after **b** clustering and **c** cancellation of non-representatives; and **d** result after and reduction

dinates are also updated when suppressing adjacent boxes to improve localization accuracy. In terms of time cost, the worst-case network management complexity is $O(N^2)$, where N is the number of boxes. Therefore, when the number of boxes is large, the time cost becomes remarkably high. This problem is exacerbated in crowded scene object detection, as thousands of boxes are created in the RPN of the Faster R-CNN, and the number of candidate boxes can even be over 10,000, especially now that detection of high-resolution images is becoming more popular.

The map-reduce approach is based on [21], which is a base of our implementation, and we present a few enhancements to it here. To begin with, the [21] technique demands a sorted input, but our method takes inputs in any order. Second, [21] violates the non-transitive characteristic. We begin with the method described in [21]. The procedure is depicted in Fig. 2 [21]. Assume that each object class has a bit matrix M of size $N \times N$. The matrix is described as follows:

$$M_{i,j} = \begin{cases} 1 \text{ if } i > j \text{ and } IOU_{i,j} > threashold \\ 0 \text{ otherwise} \end{cases}$$

$M_{i,j}$ is a symmetric matrix, as shown in Fig. 2, since $IOU_{(i,j)} = IOU_{(j,i)}$. All of the 1 components in the lower triangle are marked with as -1. This is the stage of the map. Row by row, the matrix is reduced: if a row of the matrix has a -1 element, output 1, else output 0.

This method parallelizes the NMS algorithm and achieves remarkable achievement, but their tests are incomplete. In the case of different numbers of boxes, this GPU-based algorithm also has different acceleration effects. Our research will further test this algorithm, and on this basis, we make a little improvement to make it more suitable for object detection.

2.5 Faster R-CNN with GPU

Faster R-CNN [14] is the research object of another study [8]. Their study also employs GPU and CUDA to produce Faster R-CNN and has obtained excellent results, demonstrating that CUDA can be used to accelerate an existing object detection framework. Their research, on the other hand, may be better. However, there is still much space for improvement. Furthermore, they only examined the detection speed in their study and did not assess the accuracy. Because accuracy is such a crucial metric in object detection, their test findings are incomplete, and there is still much space for improvement. Their research has provided me with ideas for my study. Because accuracy is such a crucial metric in object detection, their test findings are incomplete, and there is still much space for improvement. Their research has provided us with ideas for our study.

3 Our Implementation and Optimization

In this chapter, we will first discuss how SSD512 is constructed using CUDA and how to implement and improve the NMS algorithm using CUDA. Detailed results are discussed in the next chapter.

3.1 Some Problems of Current Research

In SSD512 [19], we divide it into four parts: pre-processing layer, feature extraction layer, proposal layer and post-processing layer. We use [16] as the test object. In our test, the most time-consuming part is feature extraction layer, which accounts for 55.38%, yet the rest also accounts for nearly half of the whole. However, in the original SSD512, only feature extraction layer is calculated on GPU. The remaining three parts are all computed on the CPU. Therefore, we propose that these three parts be put on GPU to accelerate the computation. Later we will introduce our implementation method on GPU.

3.2 Pre-Processing and Proposal Layer

The preprocessing resizes the input images and removes the average RGB values. Each output pixel is given its own thread. The number of threads is equal to the number of pixels in the output. Each thread is executed in parallel on GPU. In this way, we can better preprocess images on the GPU. The pre-processing operation still accounts for a small part of the whole framework. After the pre-processing, there is

the feature extraction layer, we rewrite it with cuDNN. Another part is the proposal layer, and softmax, which is a major function in this layer, has been accelerated by using GPU in Algorithm 1, achieving a great acceleration effect.

Algorithm 1: Softmax with GPU

Input: fea_in, class_num, height, channel
Output: fea_out

```
1  i ← threadIdx.x + blockIdx.x * blockDim.x;
2  j ← threadIdx.y + blockIdx.y * blockDim.y;
3  k ← blockIdx.z * class_num;
4  if k < channel then
5  |  k_iter ← k; step ← height * channel
6  |  for k_iter < (k + class_num); do
7  |  |  coef += expf(fea_in[i * step + j * channel + k_iter])
8  |  end
9  |  coef ← 1/coef;
10 |  for k_iter < (k + class_num); do
11 |  |  fea_in[i * step + j * channel + k_iter] ←
   |  |  expf(fea[i * step + j * channel + k_iter]) * coef;
12 |  end
13 end
14 fea_out ← fea_in;
```

3.3 Post Processing

Based on the algorithm of [21], we make the main improvement in the post-processing is NMS. Our algorithm is a bit improved on the basis of [21], which makes it more suitable for object detection and can bring faster speed. NMS's purpose is to extract a single, excellent representative from each clustered candidate object detection. Therefore, NMS shares the same issues with the conventional clustering problem. It generally consists of two fundamental operations: (1) determining which cluster each detection belongs to, and (2) determining a representative for each cluster. An NMS kernel must disclose a parallelization pattern in which each processing thread independently assesses the overlapping between two specified bounding boxes in order to use the underlying architecture of general-purpose GPUs. The goal is to prevent data dependencies that serialize calculations to the greatest extent possible, overcoming the scalability limits imposed by the traditional iterative clustering procedure. Our solution solves this problem by employing a map/reduce parallelization pattern that uses a boolean matrix to encode potential item detections as well as calculate cluster representatives.

Data flow of the our method. Each dark grey box represents a block and light grey box (in Fig. 2) a thread. The numbers within the thread are the default boxes'

Algorithm 2: Our NMS method with GPU

Input: Scores and default Boxes
Output: Index Output

```
__shared__ score[NMS_MAX] ← scores[threadIdx.x];
__shared__ bit[NMS_MAX] ← 1;
thread_score ← score[threadIdx.x];
thread_index ← index[threadIdx.x];
//Sort the default boxes by their scores;
default_box1 ← default_box[thread_index];
__syncthreads();
// Compute IoU and eliminate boxes;
for i : 1 → NMS_MAX; do
    default_boxes2 ← default_box[score[i]];
    IoU← computeIoU(default_box1, default_box2);
    if IoU > IoU_threshold then
        bit[threadIdx.x] ← −1;
    end
    __syncthreads();
end
indexOut[threadIdx.x] ← bit[threadIdx.x] ∗ threadindex
```

score-index pairs. The default boxes are sorted by their scores. The sorted result is placed in the shared memory. Each thread fetches the default box with the highest score in the memory and computes the IoU for elimination. To use the Algorithm 2, boxes have to be sorted according to their scores. Similarly, we implemented the sorting algorithm on the GPU and improved mergesort's algorithm to parallelize the processing [7]. We also eliminated boxes with lower scores, because the number of boxes left affects the accuracy of detection, we will discuss the specific data in the next chapter. There are three stages to the parallel merge sort. We partition the input data into 'p' equal-sized pieces in the first step. The second step involves sorting all 'p' blocks using 'p' thread blocks. Sorted blocks are integrated into the final sequence in the last phase. Let's look at an example to better grasp the notion of parallel merge sort. In the first phase, assign each thread to a number in the unsorted array [7]. Figure 3 shows an example of parallel merge sort with two blocks and four threads per block.

In Fig. 3, the blocks are sorted using the $sortBlocks()$ function. To do so, each block is first compared to the element next to it, and then the elements are sorted. So the group is made up of four items, and the third step repeats until the block is full with sorted elements. The blocks are merged using the $mergeBlocks()$ function. We combine the blocks to create a bigger block, but we arrange the items in the resulting array such that they are ordered. As the size of the block doubles, this function is called until there is only one block left.

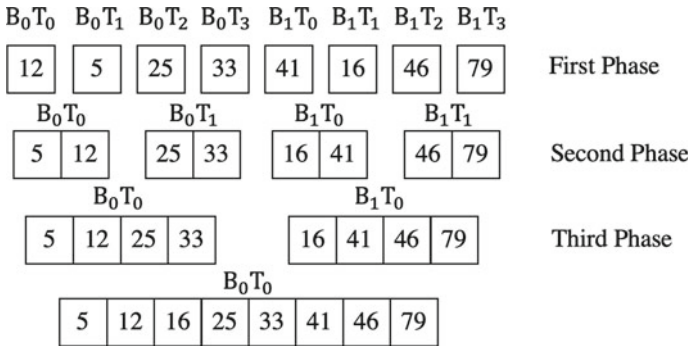


Fig. 3 Merge sort with CUDA. This is an example of a CUDA version of Merge Sort

Table 1 This is our test environment

CPU	Intel(R) Xeon(R) CPU E5-2678 v3
GPU	GeForce GTX 1080 Ti
OS	Ubuntu 18.04
CUDA	11.0
cuDNN	8.0
OpenCV	4.2

4 Result

In this chapter, we will present the benchmark results of our implementation. The result of the SSD512 will be first given, then we will discuss the performance of the NMS algorithm. Our test environment summarized in Table 1.

This time our comparison object is [16], and the data set used for our tests is PASCAL VOC2007 [5], which is a classic data set with 20 categories of objects to be identified. I will conduct various analyses of correctness and speed later. First of all, SSD512 consists of four parts: pre-processing, feature extraction, proposal layer, and post-processing. The time ratio among them is not reported yet, which is measured on our environment and shown in Fig. 4. The most time-consuming part is feature extraction, which accounts for 55.38% of the total framework and 51.48% of VGG16. The second most time-consuming part is the post-processing, where we put most of our effort. In the original paper [19], they only tested the overall FPS, but didn't do any testing for the time ratio of each part.

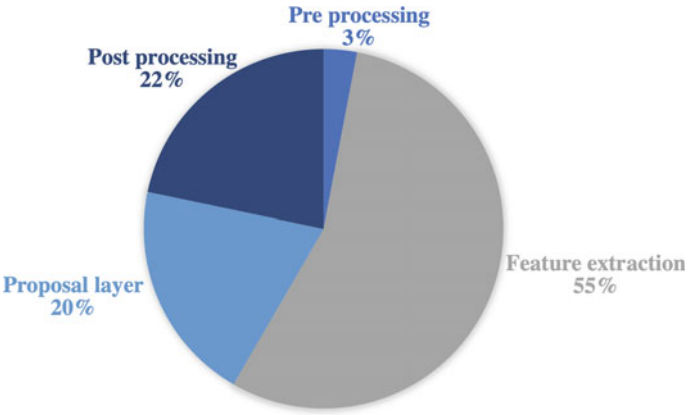


Fig. 4 Pytorch version of SSD512 time ratio. This is the overall time ratio diagram of SSD512, where the part that takes the most time is the part of feature extraction

4.1 Accuracy

To confirm the validity of our rebuilt SSD512 network, the accuracy is compared with the original work [19]. Precision and recall are two criteria used to assess accuracy. Each output bounding box's accuracy may be classified into four categories: true positive, true negative, false positive, and false negative. Precision is defined as follows:

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

The number of true positives and false positives is denoted by TP and FP, respectively. The recall is defined as:

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

where FN denotes the number of false negatives. The precision indicates how many default boxes are used to designate the proper placements of target objects among the outputs, i.e. the output bounding boxes' correctness. The recall refers to the number of target items that can be caught in the final output. Because items in datasets are separated into many classes, another metric called "mean average precision" (mAP) is established to reflect the overall precision:

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} Precision(c)$$

Table 2 The mAP comparison of our SSD512 with the original. Experiments are run with PASCAL VOC 2007 dataset. "Proposals" means the number of default boxes generated in the final output of SSD512. The right column shows that in different proposals, our accelerated SSD512 is almost as accurate as the original one

Model	Proposals	mAP(%)
Original SSD512	200	75.6
	400	76.4
	1000	76.8
	4000	77.1
Our accelerated SSD512	200	75.2
	400	76.1
	1000	76.4
	4000	76.7

where $|classes|$ denotes the number of classes. We conducted an experiment to evaluate the SSD512's mAP. The author of the study provides the performance of the original SSD512. The experiment uses the same PASCAL VOC 2007 [5] dataset as the SSD512 publication. The results are compared with the reported value in Table 2. The number of default boxes created in the final output of SSD512 is referred to as "proposals".

As shown in the Table 2, our mAP is almost equal to the original SSD512, and the error between the two is within a reasonable range. In addition, we can see from the table that as the number of boxes increases, the mAP increases gradually, we observed that, however, mAP is not improved if the number of boxes exceeds over 4000, and eventually starts to decrease.

4.2 Speed

In this section, we will present our speed test results.

As shown in the Fig. 5, significant improvement is observed for the proposal layer and the post-processing layer. We rewrote the feature extraction layer, which also made it about 9% faster than the original part. As shown in the figure, significant improvement is observed for the proposal layer and the post-processing layer. The implementation of the proposal layer is 28.44% faster than the original version, and the post-processing layer is 52.47% faster. Such a fast effect could not be achieved without our improved algorithm and the full utilization of GPU. Although speed-up in the feature extraction layer is not very obvious, it also plays an indispensable role in the acceleration of the whole SSD512 framework. On the whole, we are 22.5% faster than the original version, which further proves that it is feasible to use GPU to accelerate the current object detection framework.

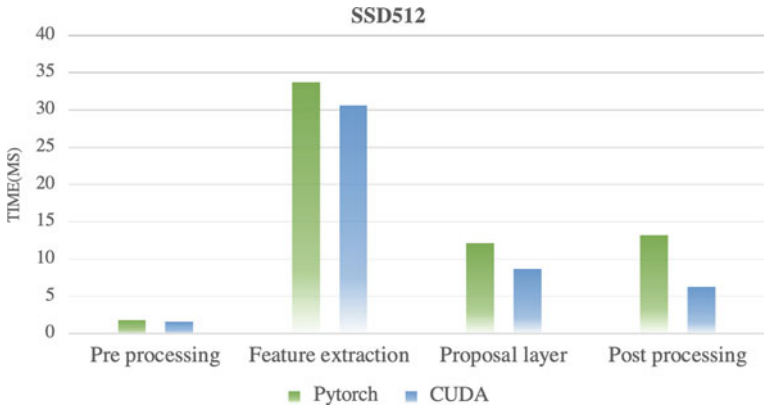


Fig. 5 Execution time Comparison Result. This is a comparison of our implementation results with the original SSD512. All parts of our implementation are faster than the original SSD512

Table 3 Execution time Of NMS and sorting. We set the threshold 0.3. The left column shows the number of boxes for the proposal and the number of boxes left after sorting. The middle column shows the results of the tests on the CPU and the right column shows the results of the GPU tests

Proposal/Boxes	CPU(ms)	GPU(ms)	Speed-Up
200/24564	5.82	3.47	×1.68
400/24564	6.37	3.58	×1.78
1000/24564	7.46	4.01	×1.86
4000/24564	13.13	6.47	×2.03

Table 4 Execution time of Different NMS. This table is our test of various NMS algorithms, and it shows that our method is the fastest

Proposal	200	400	1000	4000	20000
Faster python	0.67	1.32	2.17	8.06	78.61
Map-reduce	0.28	0.54	1.29	5.69	42.67
Our Method	0.08	0.13	0.66	3.28	28.75

Next we will focus at the NMS algorithm(with sort). In our test, there are 24,564 boxes, and the threshold is set to 0.3. And We eliminate the low-scoring boxes in the sorting process. Then we sorted the rest, taking the first 200/400/1000/4000 boxes. There are different numbers of boxes, so the test results are different. In the previous overall test, we used data of 4000 default boxes. In the case of 4000 boxes, this is the most time-consuming, but the relative mAP is also the highest (Table 3).

Our implementation on GPU is at least 1.6 times faster than on CPU and becomes faster as the number of boxes is increased to achieve higher mAP. A comparison with existing implementations is shown in Table 4. In this Table, only the NMS timing (without sort) is shown. As can be seen from Table 4, when the number of boxes is

Table 5 CPU and GPU performances of NMS with different batch size. The method can guarantee a speed-up ratio of at least 10 with batch size 1 and 400 proposals. And as the batch size gradually increases, the acceleration effect becomes more and more obvious

Batch Size	CPU(ms)	GPU(ms)
1	1.32	0.129
2	3.16	0.147
4	7.51	0.162
8	16.68	0.193
16	35.93	0.285

gradually increasing, our method is becomes more advantageous, but the speed of growth is gradually slowing down. This is because the number of GPU threads is insufficient, and we need to wait for the previous threads to complete the task and release it.

4.2.1 Different Batch Size

NMS algorithms of different batch sizes remove related boxes for different types of boxes by using NMS, which will greatly increase the speed. The different Batch size method is used in the NMS algorithm, which can be used to separate different kinds of boxes. This method is achieved by adding an offset to boxes of different types. However, the more categories, the more memory will be used, so the speed will be reduced accordingly Table 5.

With batch size 1, the approach can ensure a speed-up ratio of at least 10. The GPU method's time does not scale linearly with batch size; in fact, when the batch size is less than 16, the time is roughly the same. The reason for this is that when the batch size is less than 16, the overhead of the kernel launch consumes the majority of the time. As the batch size grows, thread-level parallelism grows as more threads are started, allowing more calculations to be completed in the same amount of time. When the batch size is equal to or more than 16, the number of threads launched may exceed the number of threads that a GPU can operate at once, causing a longer execution time. When the NMS is operating on the CPU, each image takes 1.32ms to process when the batch size is 1 and the number of proposals is 400. Our approach in Table 3 takes 47.17 ms to perform an image inference where the batch size is 1 and the number of proposals is 4000, if the NMS runs on the GPU, it will take up 13.72% of the whole time. The NMS will finish in roughly 3ms when utilizing the GPU, which may be disregarded throughout the pipeline. This is a significant performance speedup.

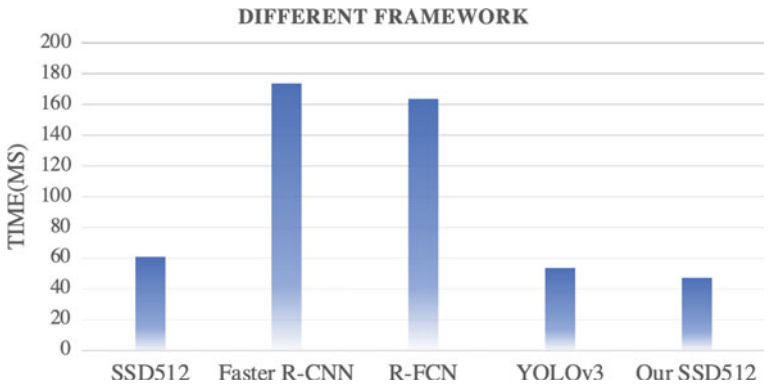


Fig. 6 Execution time Comparison Result with Other Framework. This is the speed of our accelerated SSD512 compared to several other target detection frameworks

4.2.2 Other Framework

Although it is difficult to make a fair comparison, we have tried to compare our implementation with other frameworks. There is no simple solution to the question of which model is the best. We make decisions in real-world applications to strike a balance between accuracy and speed. Aside from detector kinds, we must be mindful of other options that have an influence on performance such as Feature extractors (VGG16, ResNet, Inception, MobileNet), Non-max suppression IoU threshold, Boundary box encoding, and so on. In Fig. 6, our speed test results for different kinds of object detection frameworks. It can be seen that the overall speed of the two-stage framework [4, 14] is slower than that of the one-stage framework [23], while the two-stage frameworks obtain higher accuracy. Anyway, as far as the detection speed is concerned, our implementation is fastest among them.

5 Conclusions

We used GPU to accelerate SSD512. Our speed-up approaches are also applicable to other significant detection networks such as R-FCN, YOLO, and M2det [28], since we obtained a speed-up of the common fundamental computation of detection networks. With the default parameters of the Pytorch version [16] whose CNN layer is VGG16, we tested the speed-up of SSD512. Convolution and complete connection layers still require a long time to process, as seen in Fig. 5. If we apply our ideas to a network that spends less time computing convolution, we anticipate seeing a more substantial speedup. Our test is run on a pretty strong GPU, but if we want to commercialize our technology, we'll need to consider GPUs with lower power

consumption, which has lesser performance but may be used in a wider range of applications, such as vehicles, unmanned aerial vehicles, and so on.

Acknowledgements This work was partly supported by JSPS KAKENHI Grant Number 20H04165.

References

1. Nvidia cuda home page. https://developer.nvidia.com/zh-cn/cuda-toolkit (2017)
2. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569 (2017)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
4. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html (2007)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
7. Faujdar, N., Ghrera, S.P.: Performance evaluation of merge and quick sort using gpu computing with cuda. Int. J. Appl. Eng. Res. **10**(18) (2015)
8. Fukagai, T., Maeda, K., Tanabe, S., Shirahata, K., Tomita, Y., Ike, A., Nakagawa, A.: Speed-up of object detection neural network with gpu. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 301–305. IEEE (2018)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
11. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2888–2897 (2019)
12. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4507–4515 (2017)
13. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7310–7311 (2017)
14. Jiang, H., Learned-Miller, E.: Face detection with the faster r-cnn. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 650–657. IEEE (2017)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, p. 25 (2012)
16. Li, C.: High quality, fast, modular reference implementation of SSD in PyTorch. https://github.com/lufficc/SSD (2018)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Lawrence Zitnick, C.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)
20. Mehta, S., Misra, A., Singhal, A., Kumar, P., Mittal, A.: A high-performance parallel implementation of sum of absolute differences algorithm for motion estimation using CUDA. In: HiPC Conf, p. 6 (2010)
21. Oro, D., Fernández, C., Martorell, X., Hernando, J.: Work-efficient parallel non-maximum suppression for embedded GPU architectures. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1026–1030. IEEE (2016)
22. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. Preprint at arXiv:1612.08242 (2016)
23. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. Preprint at arXiv:1804.02767 (2018)
24. Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
25. Sachetto Oliveira, R., Rocha, B.M., Amorim, R.M., Campos, F.O., Meira, W., Toledo, E.M., Santos, R.W.D.: Comparing CUDA, OpenCL and OpenGl implementations of the cardiac monodomain equations. In: International Conference on Parallel Processing and Applied Mathematics, pp. 111–120. Springer (2011)
26. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769 (2016)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint at arXiv:1409.1556 (2014)
28. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9259–9266 (2019)

An Empirical Study on the Economic Factors Affecting on the Export of Defense Industry Using Hofstede's Culture Dimension Theory

Taeyeon Kim, Dongcheol Kim, Jaehwan Kwon, and Gwangyong Gim

Abstract The defense industry export of Korea has been consistently increased since the Defense Acquisition Program Administration (DAPA) was established in 2006; it has positive impacts on the domestic economy. Nevertheless, as competition between countries has gradually intensified in the global defense industry market, efforts to occupy the market for defense exports also have been constantly demanded. Therefore, this thesis proposes the necessity of considering the cultural factors of the countries listed to export in order to establish a defense export marketing strategy effectively. To support the proposal, the defense export marketing strategies and purchasing models of the other counties, that are not clearly established at present, was proposed and partially verified in this thesis by conducting statistical analysis using data related to defense, national, and military size of each country in the world. Through this study, it was confirmed that cultural factors are related to indirect influences on defense exports. Therefore, further discussion is needed to consider cultural factors in developing future defense export marketing strategies.

Keywords DAPA · Defense industry · Marketing strategy · Cultural factors · Indirect influences

T. Kim · G. Gim (✉)
Department of Business Administration, Soongsil University Seoul, Seoul, South Korea
e-mail: gygim@ssu.ac.kr

T. Kim
e-mail: tyk1708@gmail.com

D. Kim · J. Kwon
Department of IT Policy and Management, Soongsil Univ. Seoul, Seoul, South Korea
e-mail: onodckim@ssu.ac.kr

J. Kwon
e-mail: wcdma0225@gmail.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_8

1 Introduction

Nowadays, various factors such as society and culture in addition to economic factors affect exports of the defense industry unlike in the past when military scale and national economic power determine the flow of international relations and affect exports of the defense industry. In particular, cultural activity is used to promote friendship between countries, and furthermore, it contributes to increasing exports of the defense industry. Therefore, there is a need to consider cultural factors besides military and economic factors (defense budget, military force, and etc.) when carrying out the defense export marketing.

In an international trade environment where the export market is diversified around the world, the strategy of Korea's defense export companies needs to change from a competitive advantage strategy based on price competitiveness to an export competitive advantage strategy by enhancing export marketing capabilities. An understanding of cultural differences is essential for the segmentation of the target market, which is the core of exporting marketing, and if the culture of the importing country is different from that of the exporting country, the trade negotiation strategy has to be different [1].

The Korean government is targeting niche markets with various policies and promotional strategies to boost exports of the defense industry despite difficult circumstances. Korea's defense export figure was only 0.25 billion dollars in 2006, when the Defense Acquisition Program Administration was established, but has continued to increase and achieved an average of more than 3 billion dollars annually since 2013 through the government's export support activities and the defense industry's export efforts. The performance of defense exports is showing a noticeable increase as the government's active investment in the defense industry and direct intervention in defense sales diplomacy for the export of domestic R&D weapons systems. Therefore, it is an important time when an integrated and strategic approach by companies and government is needed to continue to expand our defense exports at a time when defense spending is being reduced globally [2, 3].

This study identified the possibility of applying cultural considerations to defense export marketing strategies in addition to existing considerations.

The study proposes a conceptual model of defense export that links defense export marketing strategies and purchasing models of importing countries; seeks the applicability of marketing strategies according to the cultural characteristics of each country.

To this end, the study analyzes whether marketing strategies can be applied differently according to cultural characteristics between countries by utilizing real-world data. Through this study, the study aims to contribute to the development of ideas for marketing strategies in consideration of the cultural factors of each country when governments and businesses establish defense export strategies.

2 Theoretical Background

2.1 Definition of Defense Industry

The defense industry is an industry engaged in the production and development of military supplies, including direct combat weapons such as weapons and ammunition, as well as general supplies such as sheath and military supplies, but usually refers to industries engaged in the production and development of critical devices that forms a national defense power such as guns, ammunition, ships, and aircraft [4].

2.2 Features of Defense Industry

The most important feature of the defense industry is the 'security industry'. National security can be defined as 'protecting all national human and physical assets and all tangible and intangible values against external threats.' There are various means and methods to protect them, but above all, diplomacy and defense must be premised [5].

The second feature of the defense industry is that it is an industry integrated with high-technology [6].

The third feature is that the defense industry is an industry with a consumer-centered market structure. The defense industry has a bilateral monopolistic market structure, with the only demand of the military, and only a single or a small number of major suppliers.

2.3 Features of Defense Industry Export

Defense industry export generally means exporting military supplies, such as defense products, or defense science and technology abroad. It has various features that are very different from exporting commercial products [7].

First, the biggest feature is that it is a "Bilateral monopoly" that does not apply to market economy logic. In other words, products (weapon systems) produced by some limited producers (defense companies) are sold to particular buyers (foreign governments). This applies both domestically and internationally [8].

Second, most of the defense industry exports are carried out as secret projects, and joint efforts by the private, government, and military are needed. For that reason, it is extremely difficult to export weapon systems by defense companies alone, so strategic support from the government is essential. Various departments, including the government as well as defense companies, can achieve defense exports through close cooperation [9, 10].

Third, defense industry export must have close trust and cooperation between the exporting and importing countries, as well as necessary to consider diplomatic

and security situations such as regional circumstances in neighboring countries surrounding the two countries. Since the weapons system is closely related to national security, neighboring countries or hostile countries should be considered [11, 12].

Finally, defense industry export is a long-term business. Unlike commercial products, the weapon system has a very long-life cycle [13]. Once the weapon system is adopted, the weapon system must be used for 30 to 40 years. Subsequent military support for maintaining operations should be considered during the use of the weapons system. Thus, the exporting and importing countries should maintain a desired relationship until the weapons system is scrapped.

2.4 Definition of Hofstede Insights

In general, cultural factors are strong influencers in consumer choice behavior and value judgment and are shown by differences in social members' preferences and ways of satisfying their needs [14]. It is impossible to define cultural factors in one way because cultural variables vary in age, gender, language, and history. Therefore, many scholars have tried to understand the cultural background of a country, and among them, cross-cultural scholar Hofstede proposed ways to quantify cultural factors through six representative cultural indicators (Index) [15].

Dr. Hofstede's Cultural Dimension Theory is defined as a total of six variables and has been continuously studied to date. Hofstede's theory has limitations that it is not significant at the level of local research and organization within the country but is known to be very useful in identifying differences between cultures based on empirical research [16] (Table 1).

3 Study Direction and Results

Sousa, etc. (2008) analyzed 52 papers from 1998 to 2005, and Chen, etc. meta-analyzed 124 papers from 2006 to 2014; according to their analysis, it was found that the culture of the other country had a decisive impact on international marketing [17, 18]. Therefore, the marketing strategy of defense exports needs to be approached differently from the general way. It will help to establish an export strategy if the results of the identification and correlation analysis of factors affecting defense exports and the empirical analysis of how they affect export performance are referenced [19, 20].

This study presented a conceptual model that distinguishes factors affecting a country's defense exports (the amount of defense imports) into two categories (independent variables and regulatory variables).

Table 1 Hofstede's influencing factor

Influencing factor	Meaning	Low	High	Main country
Power distance	The concept of how much power is distributed to members of a society or organization	Equality orientation	Hierarchical orientation	China
Individualistic	Believing that personal profits should be prioritized over collective profits	Group-centered	Individual-centered	USA, UK
Uncertainty avoidance	Cultural characteristics of how people react to dangerous situations	Risk taking	Risk avoidance	Japan, China
Masculinity	Tendency to emphasize and value the conventional role of sex	Masculinity non-oriented	Masculinity oriented	Japan, France
Long-term orientation	Having a practical and future-oriented perspective	Short-term orientation	Long-term orientation	Republic of Korea, Japan
Indulgence	Attitudes toward the enjoyment of human life and the satisfaction of possession	Suppression of pleasure	Pleasure oriented	South America

The direct defense-related factors (national defense budget, GDP, military force, mid-term defense budget, total population, and etc.) and the size of defense imports are expected to be correlated. However, it is difficult to develop into a marketing strategy using these influential factors. For instance, even if we understand the correlation that the bigger the military force, the larger the amount of defense imports, it is difficult to use it for marketing strategies [21].

Therefore, we set up a hypothesis that depending on the difference in cultural characteristics, it is likely to affect the size of defense imports. The size of defense imports is expected to be higher in cultures that value national security and avoid risks, and in cultures that prefer a long-term perspective. Thus, after identifying the cultural characteristics that influence defense imports, it is necessary to establish a marketing strategy for defense exports in consideration of cultural characteristics unlike direct influencing factors [22, 23].

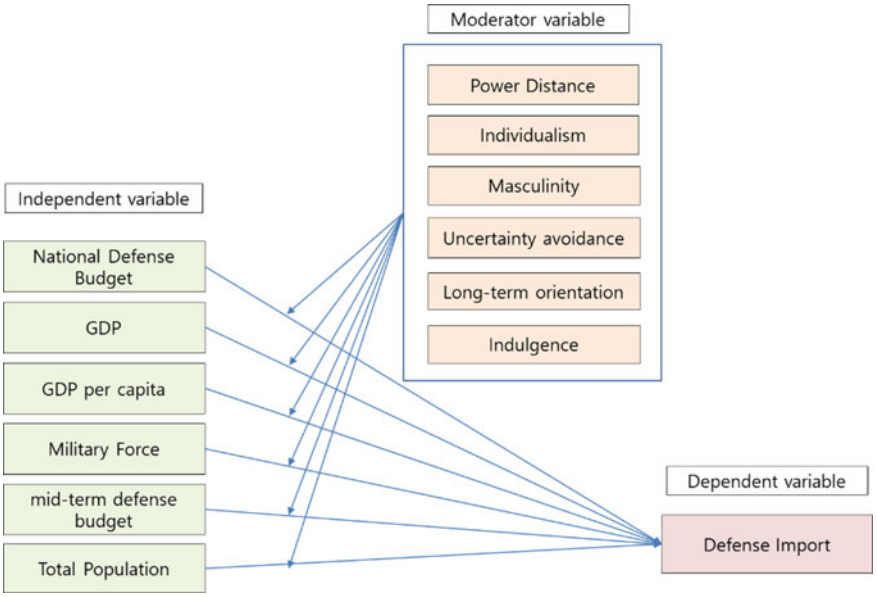


Fig. 1 Research model

3.1 Research Model

The research model for this study is shown below as (Fig. 1). Based on the theoretical background, in order to conduct an empirical analysis on the influence of economic factors of defense industry and Hofstede's cultural dimension on defense imports, the economic factors of the defense industry are set as independent variables, the amount of defense imports as a dependent variable, and the cultural dimension of Hofstede as the moderator variable; the following research model was established.

3.2 Selection of Variables

3.2.1 Selection of Independent Variables, Dependent Variables, and Moderator Variables

The research model of this study is a model to analyze the impact of defense industry's economic factors on defense exports (imports). Therefore, the economic factors of the defense industry are set as independent variables, and defense import is set as dependent variables [24].

In this study, independent variables and dependent variables are selected in Ha Kwang-ryong's (2018) paper, which demonstrated the determinant of arms trade in

Table 2 Independent variable, moderator variable, and dependent variable

Independent variable	Moderator variable	Dependent variable
National defense budget	Power distance	Defense import
GDP	Individualism	
GDP per capita	Masculinity	
Military force	Uncertainty avoidance	
Mid-term defense budget	Long-term orientation	
Total Population	Indulgence	

the global arms market, and the moderator variables are selected from samples of Hofstede's cultural dimension.

Hofstede's cultural dimension theory is defined as a total of six variables and has been continuously studied to date. Hofstede's theory has limitations that it is not significant at the level of local research and organization within the country but is known to be very useful in identifying differences between cultures based on empirical research.

These independent variables, moderator variables, and dependent variables are arranged in (Table 2).

3.2.2 Operational Definition of Independent, Dependent and Moderator Variables

The operational definitions of the variables used in this study are described in (Table 3).

4 Empirical Analysis

4.1 Effects of Economic Factors in Defense Industry on Defense Export

As a result of the correlation analysis, defense import has a strong positive correlation with defense budget, GDP, military force, mid-term defense budget, and total population variables [25]. This shows that the defense budget, GDP, troops, mid-term defense budget, and the total population have a relatively large impact on defense exports and defense imports. GDP also has a strong positive correlation with defense budget, military force, mid-term defense budget, total population, and defense import. The correlation between other variables are shown in (Table 4).

Table 3 Operational definition of variables

Category	Variables	Operational definition
Independent variable	Defense budget	Budget required to organize and maintain the national army
	GDP	Market value of all final products produced within a country's borders over a period of time
	GDP per capita	GDP divided by the average population of the year
	Military force	Number of troops in major countries that have defense imports
	Mid-term defense budget	Defense budget required after a period (5–10 years)
	Total population	Total population in major countries that have defense imports
Moderator variable	Power distance	The concept of how much power is distributed to members of a society or organization
	Individualistic	Believing that personal profits should be prioritized over collective profits
	Masculinity	Tendency to emphasize and value the conventional role of sex
	Uncertainty avoidance	Cultural characteristics of how people react to dangerous situations
	Long-term orientation	Having a practical and future-oriented perspective
	Indulgence	Attitudes toward the enjoyment of human life and the satisfaction of possession
Dependent variable	Defense import	Amount of defense imports of countries with potential defense exports

The regression standardization coefficients and hypothesis verification result in (Table 4) shows that the factors influencing the defense import, which is dependent variable were defense budget, GDP per capita, military force, defense medium budget, and total population at 95% confidence level. It shows that the national defense budget has negative impacts, and the other factors have positive impacts on the defense import.

The fact that the national defense budget has negative impacts on the defense import is that countries with higher defense budgets have a negative impact on defense weapons imports because they prefer to secure defense technology through R&D in their own countries.

Table 4 Regression standardization coefficients and Hypothesis validation results

Category	Non-standardization coefficients		Standardized coefficient	t	Significance probability	Hypothesis testing
	B	Standard error	Beta			
Constant	−256.231	82.729		−3.097	0.003	
Defense budget	−0.072	0.010	−3.836	−6.951	**0.000**	**Accept**[a]
GDP	0.000	0.000	−0.170	−1.028	*0.309*	*Reject*
GDP per capita	0.006	0.002	0.078	2.355	**0.022**	**Accept**[a]
Military force	0.705	0.287	0.189	2.459	**0.017**	**Accept**[a]
Mid-term defense budget	0.068	0.009	4.100	7.240	**0.000**	**Accept**[a]
Total population	0.000	0.000	0.341	2.368	**0.022**	**Accept**[a]

[a] $P < 0.05$

4.2 *Analysis of the Moderate Effect of Hofstede Cultural Index*

The total results of hypothesis testing of the research model by moderator variables are as shown in (Table 5).

The GDP, which is rejected in the hypothesis test of independent variables, is shown to have a moderate effect on defense imports due to influence of cultural factors, and the GDP per capita among the variables adopted is analyzed to have no moderate effect. In particular, variables such as power distance, individualism, long-term orientation, and indulgence among moderator variables are proven to have a large moderate effect, with most of the R-square changes exceeding 50%.

This study intended to establish and verify a hypothesis that the Hofstede's Cultural Dimensions Index has a moderate effect in relation to the effect of economic factors in defense industry on defense import. The summary of research hypothesis validation results is as shown in (Table 6).

4.3 *Case Study of Moderator Variables Through Hypothesis Test Results*

The analysis of the factors influenced by the moderator variables through the actual examples of major importers are as shown in (Table 7).

In general, countries with high GDP (China, Korea, and etc.) have a high defense import ranking. However, some countries (Iraq, Morocco, Pakistan, Singapore,

Table 5 Correlation analysis of economic factors in defense industry

Dependent variables	Independent variables	Moderator variables	R square	R square variance	F variance	Significance probability F variance	Hypothesis testing
Defense import	Defense budget (**Accept**)	Power distance	0.858	**0.710**	264.780	0.000	**Accept**
		Individualistic	0.937	**0.745**	627.226	0.000	**Accept**
		Masculinity	0.231	0.096	6.593	0.013	**Accept**
		UAI	0.478	0.320	32.466	0.000	**Accept**
		Long term orientation	0.890	**0.702**	337.2967	0.000	**Accept**
		Indulgence	0.887	**0.737**	344.807	0.000	**Accept**
	GDP (*Reject*)	Power distance	0.927	**0.613**	447.636	0.000	**Accept**
		Individualistic	0.927	**0.550**	397.367	0.000	**Accept**
		Masculinity	0.319	0.028	2.217	0.142	*Reject*
		UAI	0.565	0.256	31.233	0.000	**Accept**
		Long term orientation	0.834	**0.514**	164.331	0.000	**Accept**
		Indulgence	0.896	**0.574**	292.112	0.000	**Accept**
	GDP per capita **(Accept)**	Power distance	0.011	0.000	0.019	0.890	*Reject*
		Individualistic	0.027	0.002	0.118	0.732	*Reject*
		Masculinity	0.035	0.016	0.870	0.355	*Reject*
		UAI	0.087	0.028	1.596	0.212	*Reject*
		Long term orientation	0.133	0.042	2.544	0.117	*Reject*
		Indulgence	0.010	0.000	0.000	0.991	*Reject*
Defense import	Military force **(Accept)**	Power distance	0.645	0.056	8.424	0.005	**Accept**
		Individualistic	0.743	0.157	32.472	0.000	**Accept**
		Masculinity	0.810	0.223	62.283	0.000	**Accept**
		UAI	0.851	0.239	85.130	0.000	**Accept**
		Long term orientation	0.779	0.187	44.862	0.000	**Accept**
		Indulgence	0.651	0.063	9.498	0.003	**Accept**
	Mid-term defense budget **(Accept)**	Power distance	0.946	**0.694**	678.245	0.000	**Accept**
		Individualistic	0.975	**0.673**	1,410.308	0.000	**Accept**
		Masculinity	0.385	0.150	12.916	0.001	**Accept**
		UAI	0.672	0.421	68.088	0.000	**Accept**

(continued)

Table 5 (continued)

Dependent variables	Independent variables	Moderator variables	R square	R square variance	F variance	Significance probability F variance	Hypothesis testing
		Long term orientation	0.946	**0.666**	650.185	0.000	**Accept**
		Indulgence	0.976	**0.698**	788.971	0.000	**Accept**
	Total population **(Accept)**	Power distance	0.928	0.030	22.171	0.000	**Accept**
		Individualistic	0.933	0.037	29.410	0.000	**Accept**
		Masculinity	0.931	0.033	25.223	0.000	**Accept**
		UAI	0.950	0.055	58.320	0.000	**Accept**
		Long term orientation	0.0965	0.067	99.670	0.000	**Accept**
		Indulgence	0.936	0.038	32.049	0.000	**Accept**

Table 6 Summary of research hypothesis validation results

No	Hypothesis	Result
1	Economic factors in the defense industry have a significant impact on defense import	
1–1	Defense budget has a significant impact on defense import	**Accept**
1–2	GDP has a significant impact on defense import	*Reject*
1–3	GDP per capita has a significant impact on defense import	**Accept**
1–4	Military force has a significant impact on defense import	**Accept**
1–5	Mid-term defense budget has a significant impact on defense import	**Accept**
1–6	Total population has a significant impact on defense import	**Accept**
2	The cultural indices play a role as a moderator in the influencing relationship between economic factors of the defense industry and defense import	
2–1	The cultural indices play a role as a moderator in the influencing relationship between defense budget and defense import	**Accept**
2–2	The cultural indices play a role as a moderator in the influencing relationship between GDP and defense import	**Partly accept**
2–3	The cultural indices play a role as a moderator in the influencing relationship between GDP per capita and defense import	*Reject*
2–4	The cultural indices play a role as a moderator in the influencing relationship between military force and defense import	**Accept**
2–5	The cultural indices play a role as a moderator in the influencing relationship between Mid-term defense budget and defense import	**Accept**
2–6	The cultural indices play a role as a moderator in the influencing relationship between total population and defense import	**Accept**

Table 7 Comparative analysis of factors affecting moderator variables

Country	GDP (ranking/number of countries)	Defense import (ranking/number of countries)	Power distance	Long term orientation
China, P.R.	13,407,400 (2/57)	12,667 (1/57)	80	87
Korea, South	1,619,420 (11/57)	1317 (2/57)	60	100
Pakistan	312,570 (33/57)	777 (3/57)	55	50
Turkey	766,428 (16/57)	685 (5/57)	66	46
Iraq	226,070 (39/57)	596 (7/57)	95	25
Viet Nam	241,272 (37/57)	546 (9/57)	70	57
Singapore	361,109 (28/57)	510 (12/57)	74	72
Morocco	118,309 (43/57)	387 (14/57)	70	24

Turkey, and Vietnam) have high defense import rankings despite their low GDP. In both cases, the countries have high cultural indices such as the country's power distance and long-term orientation. Therefore, cultural factors of the country are analyzed to affect defense import.

5 Conclusion and Future Research

This study is to derive direct factors affecting defense exports through real data and to find and analyze factors that indirectly affect defense exports.

Specifically, it is a practical analysis of the impact of factors affecting defense exports on major national's amount of defense imports. In accordance with the proposed research procedures, the study first identified the factors influencing the defense sector, defined the amount of defense imports, and conducted a test on the hypothesis after presenting a research hypothesis on the causal relationship between direct & indirect influencing factors and the amount of defense imports.

As the results of a study, the direct factors (national defense budget, GDP per capita, military force, Mid-term defense budget, and total population) and indirect factors (individualistic, power distance, uncertainty avoidance, Masculinity, long-term orientation, and indulgence) that influence the amount of defense imports were presented, and it was analyzed that it was reasonable to define the amount of defense imports as an index of exportable country.

The major achievements and contributions of this study are as follows:

First, the study established the research and analysis procedures for factors affecting defense exports.

Second, the study empirically verified through the Hofstede Cultural Dimension Index that not only economic factors in the defense industry but also indirect factors such as the cultural characteristics of each country can affect defense exports by the empirical analysis.

Third, the study presents a research model that can effectively analyze the causal relationship between economic factors in the defense sector, Hofstede's cultural dimension factors, and amounts of defense imports.

Fourth, the study provides a basis for verifying hypotheses based on accurate figures with empirical analysis through real-world data breaking away from the qualitative survey method.

It is true that Korea's defense exports have increased rapidly since the opening of the DAPA (Defense Acquisition Program Administration) due to environmental analysis of purchasing countries and active efforts. However, defense exports are greatly influenced by the global crisis and changes in the global defense environment. Korea is currently in the ranks of emerging defense export economies, and export markets are limited due to the nature of the defense industry. Therefore, the effective marketing strategy is needed to maintain steady defense exports.

The marketing steps can be sorted into 3 stages and approached differently in order to establish the marketing strategies.

The initial stage (stage 1) is to explore countries that are capable of defense import, the expansion stage (stage 2) is to expand defense exports with countries that started importing defense, and the final stage (stage 3) is the continuous stage of defense exports. This study identified that cultural differences affect defense industry exports especially in the early stages of exploring countries that are capable of importing defense before establishing a marketing strategy.

This paper is meaningful in that it attempted an empirical analysis through real-world data on the factors necessary to establish strategies for survival in the global defense market. Following the research methods and results of this study, the future research directions are presented as followed.

First, detailed and diversified identification of the factors affecting defense exports is required. In this study, the empirical analysis of socio-cultural factors was conducted for the first time. However, it is true that the analysis of the direct factors affecting defense exports is insufficient. For instance, it is necessary to include and analyze the World Peace Index or the Corruption Index, and etc. as factors.

Second, various studies on variables that affect defense exports are needed. The results of this study show that economic and cultural factors affect the defense industry directly and indirectly. The variables in this study were demonstrated based on the defense-related indices of major defense export countries and the cultural dimension indices of Hofstede-considering the importance of the effects of these variables on defense import, research should be done considering time-series analyses or associations between variables.

Third, research results need to be harmonized with actual policy, institutional establishment, decision-making system, and feedback. This naturally establishes a defense export marketing strategy and can lead to actual defense export performance.

References

1. Chung, Y.K.: The relationship between export marketing negotiation and national cultures in Hofstede and Hall's perspectives. IASR **11**(2), 253–290 (2007)
2. DTAQ: 2019 Global Defense Market Yearbook (2019)
3. Akerman, A., Seim, A.L.: The global arms trade network 1950–2007. J. Comp. Econ. **42**(3), 535–551 (2014)
4. DAPA: Defense Terminology Dictionary, 214 (2013)
5. Kim, Y.-N.: A Study on activating strategic exports of defense industry: focused on government subsidy policy and system reform, Ph.D. paper at Kyungnam Univ. (2011)
6. Hwang, I.B.: Market analysis and support system for export revitalization, MD paper SKG Univ. (2009)
7. KOTRA: Defense export comprehensive guidebook, 4. (2019)
8. Ha, K.Y.: Determinants of arms trade in the global arms market: focusing on the characteristics and interrelation of exporting and importing countries. KATIS **23**(1), 25–54 (2018)
9. Kil, B.-O.: South Korea's refinement plans of supporting institutions for the export revitalization of the defense industries. J. Korean Assoc. Def. Indus-Try Stud. **26**(2), 1–16 (2019)
10. Kim, S.Y.: A study on the development of defense R&D for improving the competitiveness of defense industry export, defense & technology, 392, KDIA. (2012)
11. Kim, C.M., Jang, W.J: Recent trends in defense industry exports and future challenges, KIET (2012)
12. An, Y.S.: A study on the regulations for implementation management of export contracts between governments, KIET (2017)
13. Lee, J.Y.: The performance and growth factors of Korea's defense industry export. Def. Technol. **428**, KDIA, 44–55 (2014)
14. Hye, J.K., Jae, S.L.: Individualistic/collectivistic orientations and organizational citizenship behavior: moderating effects of affective commitment and turn-over intention. JHRMR **19**(1), 47–69 (2012)
15. Hofstede, G.: Culture's Consequences: International Differences in Work-related Values, Beverly Hills. Sage Publications, Calif (1980)
16. Hofstede, G.: Culture and Organizations: Software of the Mind. Mcgraw Hill, London, England (1991)
17. Sung, M.J., Sung, J.P., Yoon, C.C.: An Analysis of the impact factors on the export performance of the defense industry: the analysis of correlations using key statistical indicators in OECD countries. Def. Technol., KDIA, 104–119 (2018)
18. Sousa, C.M.P., Martinez-Lopez, F.J., Coelho, F.: A review of the research in the literature between 1998 and 2005. Int. J. Mark. **10**(4), 343–374 (2008)
19. Moon, J.-Y., Kwon, H.-C., Kim, D.-Y.: Estimation of defense industry export function for defense industry promotion plan. J. Korean Assoc. Def. Ind. Stud. **26**(2), KADIS, 67–80 (2019)
20. Park, W.J.: The reality and development of Korea's defense export control system. Def. Technol., KDIA, 38–46 (2011)
21. Jang, W.J.: Top 10 promising countries for 2014 KIET defense exports, KIET (2014)
22. Branton, S. L.: Foreign policy in transition human right, democracy and U. S. Arms Exports. Int. Stud. Q. **49**, 647–667 (2005)
23. Eriskson, J.: L: Market imperative meets normative power: human rights and European arms transfer policy. Eur. J. Int. Rel. **19**(2), 209–234 (2011)
24. Chen, J., Sousa, C.M.P., He, X.: The determinants of export performance: a review of the literature 2006–2014. Int. Mark. Rev. **33**(5), 626–670 (2016)
25. Jane's IHS: Defense Budgets, Jane's IHS (2018)

A Study on the Application of Blockchain Technology in Non-governmental Organizations

Minwoo Lee, Saeyeon Lee, Heewon Lee, and Gwangyong Gim

Abstract Looking at the market trend of Non-Profit Organizations (NGOs), the number of non-profit private organizations is steadily increasing due to the rapid growth of civil society and as the importance of public interest activities of private organizations are highlighted. However, the size of domestic donations has been in decline since its peak in 2013. Transparency and reliability are important for donation organizations' selection criteria, and the most necessary thing to spread the donation culture is to strengthen the transparency of fundraising organizations. Blockchain technology is well known for Bitcoin, but its true value comes from that fact that it can secure transparency and reliability at low cost. In addition, it is a technology that has an advantage in strengthening security, and has recently attracted attention in many industries. Blockchain technology is a key platform in the new era that can dramatically reduce transaction costs and ensure transparency in institutions and donations of non-profit organizations required by sponsors or those who want to sponsor. It is a technology optimized for NGOs that have to operate under IT operating costs and security at a low cost. Because blockchain technology is superior in transparency and security, companies are starting to systematically research and apply it. However, non-profit organizations currently lack a lot of research cases and have little application. This study presented the implications of data for introducing and utilizing blockchain technology to non-profit organizations (NGOs) in the future and discussed the limitations of this study and future research tasks.

M. Lee · S. Lee · H. Lee
Department of IT Policy and Management, Soongsil University Seoul, Seoul, South Korea
e-mail: andy_lee@worldvision.or.kr

S. Lee
e-mail: saeyeon9@naver.com

H. Lee
e-mail: steve.hw.lee@gmail.com

G. Gim (✉)
Department of Business Administration, Soongsil University Seoul, Seoul, South Korea
e-mail: gygim@ssu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_9

Keywords Blockchain · Donation · NGO · Non-governmental organizations

1 Introduction

Looking at the market trend of Non-Profit Organizations (NGOs), the number of non-profit private organizations is steadily increasing due to the rapid growth of civil society and as the importance of public interest activities of private organizations are highlighted. However, the size of domestic donations has been in decline.

According to Statistics Korea's social survey, Korea's donation or donation participation rate continued to grow until 2011, peaking at 36.4%, and continually declined to 25.6% in 2019 and 21.6% in 2021. According to the results of the survey on the change in donation awareness and attitude, the items that accounted for a large part of the area needed to spread the donation culture overall were enhanced transparency in the operation of donation organizations (33.3%), and convenience of donation methods (3.4%) [1].

Although the donation culture is spreading due to the COVID-19 incident, according to a survey by the Chosun Ilbo's public service session, "The Better Future," 68.6% of people replied that their choice of which donation organizations to donate to is made by transparency and reliability, and 43.3% of the respondents need transparency to spread the donation culture. Not only that, 46.4% of the respondents said they did not trust the fundraising organization for the reason for not donating [2].

Blockchain technology can secure transparency and reliability at low cost. In addition, as a technology that has strengths in strengthening security, it is receiving attention in the industrial field and playing an important role in the era of the 4th Industrial Revolution. Also, the technology has potential for development and connection.

Blockchain technology is a key platform in the new era that can dramatically reduce the transparency and transaction costs of institutions and donations required by sponsors or those who want to sponsor, and is optimized for non-profit organizations (NGOs) that have to operate IT operating costs or security at low costs.

Because blockchain technology is superior in transparency and security, companies are starting to systematically research and apply it. However, non-profit organizations currently lack a lot of research cases and have little application.

This study aims to conduct an empirical study on the acceptance of blockchain technology by non-profit organizations (NGOs).

2 Theoretical Background

2.1 The Definition of Blockchain

Blockchain was first introduced in 2008 with the advent of Bitcoin, which was invented by Satoshi Nakamoto. It was also introduced in a paper titled 'Bitcoin: Apeer-to-peer electronic cash system' published in the cryptographic technology community Maine in October 2008, and in this paper, blockchain technology was described by suggesting how to prevent double payments using P2P networks [3].

Blockchain is a security technology that checks and shares the transaction books of traders who want to use banks or financial institutions, not some, so that the transaction can be traded safely. The existing transaction method is made after confirming the previous financial transaction details by an institution called a bank or the financial sector, and the biggest difference from the blockchain is that only the minimum institution called a bank or the financial sector knows such financial transaction details. Blockchain stores records of new transactions that are generated when one block exists in another block and generates blocks at 10-min intervals. When it is confirmed that the generated block is the correct transaction, all blocks are connected to the existing block in the chain on which the transaction is recorded. All transactions in which blocks are stacked one by one and connected to form a blockchain. According to Deloitte's "Technology Trends 2022," blockchain and distributed account information technology platforms among the seven technologies fundamentally change the nature of business performance across organizational boundaries, and many companies are reimagining how to create and manage tangible and digital assets [4]. Due to the various usability of the blockchain, active introduction in the public sector as well as finance is being attempted, and it is expected to develop in various forms. It is also being used in the non-profit sector.

2.2 A Study on Technology Acceptance Theory—The Unified Theory of Acceptance and Use of Technology (UTAUT)

Many researchers in the IT field recognized the importance of the introduction of these technologies and the spread of innovation and conducted research on what prior factors lead to the introduction and spread of innovation [5]. The technology acceptance model is a model that applies new technologies or services or new products to identify variables that are affected by an individual or organization's intention to accept them [6].

Since the 1980s, it is estimated that about 50% of new capital investments have been invested in in information technology in corporate organizations [7]. In addition, acceptance of information technology is a very important issue because employees must use and accept new technologies to improve productivity within the organization [8].

There is a problem with the existing technology acceptance theory. There is a difference between the causal relationship that affects the usefulness of consumption through the acceptance of information technology by individuals and the acceptance of organizational members to information technology introduced to improve organizational productivity. In addition, with the development and convergence of technology, it has been argued that existing theories and models can explain only about 40% of the user's intention to accept information technology. There is also a limitation in that existing technology acceptance models do not sufficiently support the validity of the relationship between various variables and variables. In addition, as the level of IT increases and complex and hyper-convergence develops, a situation in which research models must be expanded is emerging [8].

Based on these problems and backgrounds, eight existing theories (TRA, TAM, MM, TPB, C-TAM-TPB, MPCU, IDT, and SCT) are integrated and supplemented. Then, the Unified Theory of Acceptance and Use of Technology, which consists of four independent variables (performance expectations, effort expectations, social impact, and facilitation conditions), four moderating variables (gender, age, experience, spontaneity), one parameter (action intention), and one dependent variable (use behavior) was presented [8].

3 Research Method

In order to predict the actual acceptance of non-profit organizations (NGOs), which are potential beneficiaries of blockchain technology, It plans to build a research model focused on acceptance as follows (Fig. 1).

3.1 Research Hypothesis

The theoretical basis for the research model on the acceptance of blockchain technology by non-profit organizations (NGOs) includes the hypothesis of 14 causal relationships and the hypothesis of two moderating effects through previous studies. In addition to the four key variables suggested in the UTAUT theory: performance expectation, effort expectation, social impact, and promotion conditions, five characteristic variables of blockchain technology expected as leading factors of performance expectation and effort expectation were additionally applied to the model. Performance expectations were viewed similarly to the perceived usefulness and effort expectations of the technology acceptance model, and how the security, availability, reliability, diversity, and economy of the blockchain influenced acceptance through performance expectations and effort expectations. Also, the hypothesis shown in Table 1 was derived through the composition of the research model.

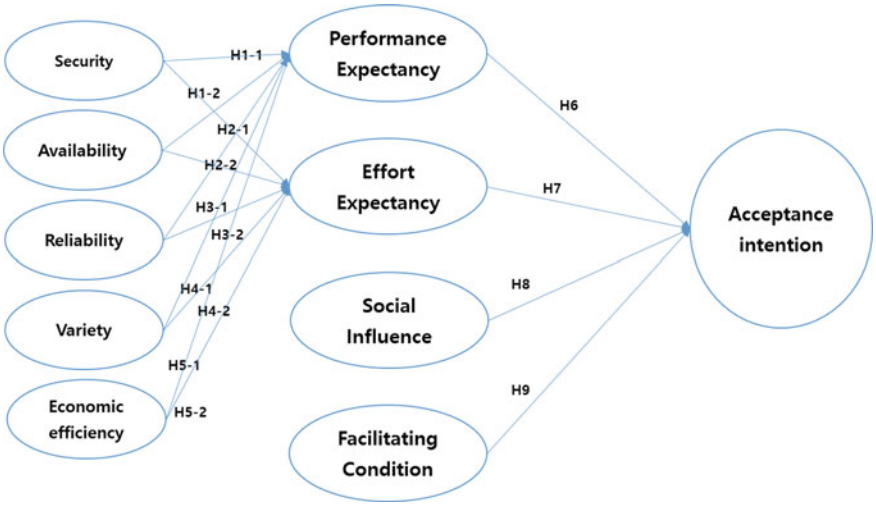


Fig. 1 Research model

3.2 Operational Definition of Variables

Operational definitions of independent variables, parameters, and dependent variables included in the research model and research hypothesis were defined and summarized.

Security is the degree to which non-profit organizations using blockchain believe that the system is safe due to intrusion and attack such as external hacking or ransomware, and that there is no possibility of forgery or tampering with data [9–13].

Availability and stability is the degree to which a non-profit organization's operating system is always available and stable to perform functions because technologies using blockchain are distributed (shared) structures [9, 10, 14–16].

Reliability and transparency are the degree to which data is accurately and transparently processed in the system of non-profit organizations that use blockchain to maintain integrity without errors in data [9, 10, 13, 17–20].

Diversity and scalability are the degree to which the system using blockchain as a non-profit organization can be used for various purposes (such as security documents) and in various fields (such as fundraising) [9, 15, 21, 22].

Economic efficiency and effectiveness are expected to be economically effective by reducing system construction and maintenance costs by introducing technology using blockchain [9, 15, 21, 22].

Table 1 Research hypothesis

H1-1	The security of blockchain technology utilization will have a positive (+) effect on performance expectations
H1-2	The security of blockchain technology utilization will have a positive (+) effect on expectations of effort
H2-1	The availability of blockchain technology utilization will have a positive (+) effect on performance expectations
H2-2	The availability of blockchain technology utilization will have a positive (+) effect on expectations of effort
H3-1	The reliability of blockchain technology utilization will have a positive (+) effect on performance expectations
H3-2	The reliability of blockchain technology utilization will have a positive (+) effect on expectations of effort
H4-1	The diversity of blockchain technology utilization will have a positive (+) effect on performance expectations
H4-2	The diversity of blockchain technology utilization will have a positive (+) effect on expectations of effort
H5-1	The economic feasibility of using blockchain technology will have a positive (+) effect on performance expectations
H5-2	The economic feasibility of using blockchain technology will have a positive (+) effect on expectations of effort
H6	Performance expectations will have a positive (+) effect on the acceptance of blockchain technology
H7	Expectations of effort will have a positive (+) effect on the acceptance of blockchain technology
H8	The social impact will have a positive (+) effect on the acceptance of blockchain technology
H9	The promotion conditions will have a positive (+) effect on the acceptance of blockchain technology
H10	Organizational innovation will regulate the relationship between blockchain technology acceptance

3.3 UTAUT Independent, Dependent Variables

Performance expectations are the degree to which it is expected to help improve work performance and the company grow by using blockchain technology. Expectation of effort is the extent to which blockchain technology is believed to be easy to use for non-profit organizations (NGOs). In addition, the social impact is the degree to which our organization or organization thinks that blockchain technology should be utilized and used in the overall environment of society. Facilitation conditions are organized to the extent of belief in the existence of an organization or a technical foundation that can support the use or use of blockchain technology in a non-profit organization [8, 9, 23–29].

Finally, acceptance of dependent variables indicates the degree to which blockchain technology is intended to be used or introduced by non-profit organizations [8, 9, 30].

4 Empirical Data Analysis

4.1 Data Collection

This study developed research models based on previous studies to understand the intention of non-profit organizations (NGOs) to accept the technology of blockchain, established research hypotheses, and conducted empirical analysis to verify each hypothesis.

In order to collect data necessary for analysis, a questionnaire was prepared to measure the variables of the research model. Next, an online survey was conducted on this. Currently, blockchain technology is in the early stages of introduction to non-profit organizations (NGOs) or there is a lack of commercialized services, so a survey was conducted on several non-profit organizations (NGOs) workers in anticipation of potential users and recipients. The survey period was conducted for 8 days from November 12, 2018 to November 20, 2018, and a total of 161 survey responses were collected. Among the data collected online, 155 cases of data were used for final analysis, excluding 6 copies that were missing or responded faithfully.

Demographic characteristics were identified through frequency analysis, reliability and validity analysis were conducted on measurement variables, and SmartPLS 3.0 were used for statistical data analysis. First, PLS Algorithm was performed to evaluate the measurement model. Next, Bootstrapping and Blindfolding were performed to evaluate the structural model, hypothesis verification, and mediating effect verification.

4.2 Reflective Measurement Model Fit

This study attempted to understand the causal relationship by applying the Smart PLS structural equation model to check the significant path coefficient between each variable presented in the research model. In addition, the appropriate model composition and the concentrated validity and discrimination validity of each item were confirmed.

4.2.1 Intensive Validity

As for the outer loading value, all measurement variables were higher than the reference value of 0.70 and the Average Variance Extracted (AVE) also exceeded the reference value of 0.5, so it can be seen that the concentration validity is secured (Table 2).

4.2.2 Reliability Verification of Measurement Model

According to the evaluation results, all of the values of kronbaha alpha were 0.60 or more, and rho_A(ρA) was higher than the threshold value of 0.70 or higher. All values of Composite Reliability (CR) were found to be 0.60 or more. Therefore, it seems that all the variables in the study have internal consistency reliability (Table 3).

4.2.3 Determination Feasibility Evaluation Results

Finally, discrimination validity should be verified to measure the degree to which variables and variables can be well distinguished. In order to evaluate discrimination validity, the HTMT criteria with the highest validity among Fornell-Larker criteria, cross loading, and HTMT (heterotrite-monotrite ratio) methods are shown in Table 4.

4.3 Structural Model Evaluation and Hypothesis Verification

The effect of the characteristics of the blockchain (availability, economy, diversity, security, reliability) on effort expectations, performance expectations, and social effects and promotion conditions was verified. In addition, the influence of parameters (efforts expectations, social effects, performance expectations, and conditions for promotion) on the acceptance of non-profit organizations' blockchain technology was verified. Table 5 shows the result of basic hypothesis verification.

This means that trustworthiness affects expectations of effort and performance, and diversity affects expectations of performance. Also, economic feasibility affects expectations of effort and performance, and performance expectations, social influences, and conditions for promotion affect the intention of non-profit organizations (NGOs) to accept blockchain.

Table 2 Intensive validity result

Variable	Measurement item	Outer loading	Indicator reliability	AVE
Availability	Availability 1	0.877	0.769	0.8
	Availability 2	0.915	0.837	
	Availability 3	0.848	0.719	
	Availability 4	0.935	0.874	
	Availability 5	0.896	0.803	
Economic efficiency	Economic efficiency 1	0.942	0.887	0.909
	Economic efficiency 2	0.964	0.929	
	Economic efficiency 3	0.96	0.922	
	Economic efficiency 4	0.968	0.937	
	Economic efficiency 5	0.934	0.872	
Effort expectancy	Effort expectancy 1	0.893	0.797	0.845
	Effort expectancy 2	0.941	0.885	
	Effort expectancy 3	0.941	0.885	
	Effort expectancy 4	0.899	0.808	
	Effort expectancy 5	0.92	0.846	
Variety	Variety 1	0.911	0.830	0.813
	Variety 2	0.945	0.893	
	Variety 3	0.89	0.792	
	Variety 4	0.861	0.741	
	Variety 5	0.898	0.806	
Security	Security 1	0.955	0.912	0.909
	Security 2	0.974	0.949	
	Security 3	0.958	0.918	
	Security 4	0.944	0.891	
	Security 5	0.937	0.878	
Social influence	Social influence 1	0.941	0.885	0.895
	Social influence 2	0.938	0.880	
	Social influence 3	0.939	0.882	

(continued)

Table 2 (continued)

Variable	Measurement item	Outer loading	Indicator reliability	AVE
	Social influence 4	0.954	0.910	
	Social influence 5	0.96	0.922	
Performance expectancy	Performance expectancy 1	0.927	0.859	

Table 3 Measurement model reliability analysis results

Factors	Cronbach's Alpha	rho_A	Composite Reliability
Availability	0.938	0.947	0.952
Economic efficiency	0.975	0.975	0.98
Effort expectancy	0.954	0.954	0.956
Variety	0.942	0.943	0.956
Security	0.975	0.976	0.98
Social influence	0.971	0.971	0.977
Performance expectancy	0.961	0.962	0.970
Acceptance intention	0.967	0.967	0.974
Reliability	0.929	0.930	0.955
Facilitating condition	0.908	0.918	0.931

Table 4 Evaluation results of discriminant validity based on HTMT criteria

	1	2	3	4	5	6	7	8	9	10
1										
2	0.667									
3	0.604	0.732								
4	0.47	0.438	0.37							
5	0.769	0.476	0.546	0.513						
6	0.572	0.635	0.656	0.507	0.545					
7	0.578	0.658	0.654	0.682	0.538	0.748				
8	0.558	0.52	0.615	0.623	0.661	0.791	0.769			
9	0.87	0.621	0.689	0.599	0.848	0.661	0.714	0.745		
10	0.413	0.505	0.682	0.296	0.46	0.704	0.504	0.716	0.554	

1 = Availability, 2 = Economic efficiency, 3 = Effort Expectancy, 4 = Variety, 5 = Security, 6 = Social Influence, 7 = Performance Expectancy, 8 = Acceptance intention, 9 = Reliability, 10 = Facilitating Condition

Table 5 Results of hypothesis testing

	Original sample (O)	P values	Adoption
Security → Performance expectations	−0.064	0.623	Dismiss
Security → Expectation of efforts	0.069	0.515	Dismiss
Availability → Performance expectations	−0.125	0.253	Dismiss
Availability → Expectation of effort	−0.124	0.257	Dismiss
Reliability → Performance expectations	0.424	0.003	Selection
Reliability → Expectation of effort	0.425	0.002	Selection
Diversity → Performance expectations	0.352	0	Selection
Diversity → Expectation of effort	−0.094	0.139	Dismiss
Economical economy → Performance expectations	0.349	0	Selection
Economical → Expectation of effort	0.543	0	Selection
Performance expectations → Acceptance intention	0.416	0	Selection
Effort expectations → Acceptance intention	−0.061	0.368	Dismiss
Social impact → Acceptance intention	0.283	0.022	Selection
Facilitation conditions → Acceptance intention	0.333	0	Selection

5 Conclusions

This study conducted an empirical study on the acceptance of blockchain technology by non-profit organizations (NGOs). Through previous studies, the characteristics of blockchain technology were investigated, and an analysis was conducted on how the mediating effect was applied to acceptance as a variable that had a mediating effect. Through various technology acceptance models, the model of UTAUT's study suitable for this study was found and designed. Research on the acceptance of new technology called blockchain in non-profit organizations (NGOs) has a meaning and importance of research that it has been a new attempt at blockchain acceptance intention to companies and public institutions.

Data were collected and empirically analyzed to verify each hypothesis between research models, and the following results could be derived based on the verification results.

First, among the independent variables of the characteristics of the blockchain, the hypothesis that security and availability affect performance expectations or effort expectations was rejected. It was interpreted that the blockchain is still considered to be just a cryptocurrency, Bitcoin, and lacks confidence that it is safe and safe against the risk of hacking. Also, most non-profit organizations (NGOs) workers who are familiar with the system of viewing, storing, and processing sponsors' data with servers in the center lack confidence in blockchain that allows stable use of P2P methods.

Second, among the characteristics of the blockchain, security and availability were rejected, but it was derived that reliability and economy influenced performance

expectations and effort expectations, and acceptance intention influenced the performance expectations. Unlike security, reliability was judged to have the belief that the trust in the information that would be contained within the block was accurate and that it would be stored safely without errors. Also, it was interpreted that the IT expenditure cost, which is the most controversial in non-profit organizations, and the existing high-priced IT construction cost can reduce maintenance costs along with system operation with low cost. In addition, diversity, a characteristic of blockchain, affects performance expectations, but the hypothesis of effort expectations was rejected.

Third, along with performance expectations and social influences, the hypothesis that facilitation conditions affect acceptance intention was adopted. However, the hypothesis that the expectation of parameter effort affects acceptance was rejected. Expectations for blockchain performance in the future, an opportunity for organizations to do new business and grow in the currently saturated domestic fundraising market, and arguments that blockchain should be introduced and done around them were interpreted as influencing acceptance intention. The judgment that blockchain technology can be educated or helped was also interpreted as having an effect on acceptance intention. This interpretation aligns with the results of many previous studies.

Finally, when blockchain technology is activated and applied to all fields, research that affects before acceptance and research on how much difference there is after acceptance are required, as in this study. In addition, by comparing the differences in how effective the variables adopted were, research will be able to contribute on finding its impact on the new technologies of future NGOs. Also, it will be a meaningful study to conduct a wide range of empirical studies targeting individuals, not groups or organizations.

References

1. Statistics Korea: The results of the 2021 Social Survey '(Welfare, Social Participation, Leisure, Income and Consumption, Labor)', 2021, Access:2022/04/23, URL: https://kostat.go.kr/portal/korea/kor_nw/1/1/index.board?bmode=read&aSeq=415115
2. Chosun Ilbo, The Better Future: "Public opinion of the better future distrusts fundraising organizations, and donors trust fundraising organizations.", 2021, Access: 2022/04/23, URL: https://futurechosun.com/archives/59549
3. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. 2008, Access: 2022/04/23, URL: https://bitcoin.org/bitcoin.pdf
4. Tech Trends 2022: Engineer your tech-forward future. 2021, Access: 2022/04/23, https://www2.deloitte.com/content/dam/insights/articles/US164706_Tech-trends-2022/DI_Tech-trends-2022.pdf
5. Kim, D.H.: A Study of Factors Affecting the Adoption of Cloud Computing. The Journal of Society for e-Business Studies **17**(1), 111–136 (2012)
6. Jeong, C.Y.: A Study of Factors Affecting Intention to Accept the Personal Information Protection Accreditation in Medical Centers. Doctoral Thesis. Soongsil University (2015)
7. Westland, C., Clark, T.: Global Electronic Commerce: Theory & Case Studies. Massachusetts Institute of Technology, London (2000)

8. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of in-formation technology: Toward a unified view. MIS Q. **27**(3), 425–478 (2003)
9. Kim, J.S.: A Study on Factors Affecting the Intention to Accept Blockchain Technology, Master Thesis. Soongsil University (2017)
10. Kim, J.H., Jung, M.H., Kim, J.M. and YOO, Y.S.:: Block Chain Prime, KOBIT CO., LTD (2016)
11. Moon, J.H.: A Study on the Improvement of Coupon Services in the Block Chain Based: Master Thesis. Dongkuk University (2017)
12. Song, S.H.: Transforming the Blockchain in the Age of Hyper-Connection. CLO **7**(82), 40–43 (2017)
13. Shrier, D., W. Wu, and A. Pentland.: Blockchain & Infrastructure (Identity, Data Security). MIT Connection Science. Part 3 (2016)
14. Lorenz, JT., Münstermann, B., Higginson, M., Olesen, PB., Bohlken, N., and Ric-ciardi, V.: Blockchain in insurance – opportunity or threat?. McKinsey & Company. (2016)
15. Go, Y.S., And Choi, H,S.: Changing Business Paradigm and Its Application - Focused on the Block Chain Technology. The Korean Society of Science & Art, 27, 13–29 (2017)
16. Kim, S.Y., Ahn, S.B.: A Study on Identifying Affecting Factors to Accept Blockchain System -Focused on Logistics Industry-. Korea Logistics Research Association **28**(1), 71–85 (2018)
17. Tung, F.C., Chang, S.C., Chou, C.M.: An extension of trust and TAM model with IDT in the adoption of the electronic logistics information system in HIS in the medical industry. Int. J. Med. Informatics **77**(5), 324–335 (2008)
18. Vatanasombut, B., Igbaria, M., Stylianou, A.C., Rodgers, W.: Information sys-tems continuance intention of web-based applications customers: The case of online banking". Information & Management **45**, 419–428 (2008)
19. Hong, S.P. H. In, K.H. Kim, K.J. Kim, S.M. Park, Y.J. Jung, H.J. Kang, J.E. Lee, S.J. Shim, and D.H. Hong,: "Study on the Plan for the Introduction of the Blockchain Technology in Financial Section", Financial Services Commission, (2016)
20. Seo, K.K.: Factor Analysis of the Cloud Service Adoption Intension of Korean Firms : Applying the TAM and VAM. Journal of Digital Convergence. **11**(12), 155–160 (2013)
21. Jovanovic, B., Rousseau, P.L.: General purpose technologies. Handbook of Economic Growth. **1**, 1181–1224 (2005)
22. Chung, S.H.: Legal Issues for the Introduction of Distributed Ledger Based on Blockchain Technology-Focused On the Financial Industry-. Korea Financial Law Association **13**(2), 107–138 (2016)
23. Suh, B.S.: A Study on the Factors Affecting the Intention to Adapt PMO in Public Sectors. Doctoral Thesis. Soongsil University (2013)
24. Jeon, S.H., Park, N.R., Lee, C.C.: Study on the Factors Affecting the Intention to Adopt Public Cloud Computing Service. Entrue Journal of Information Technology. **10**(2), 97–112 (2011)
25. Kim, Y.G.: The Effect of Perceived Risk and Trust on Users' Acceptance of Cloud Computing :Mobile Cloud Computing. Doctoral Thesis. Inchon University (2011)
26. Tai, Y.M., Ku, Y.C.: Will stock investors use mobile stock trading? A benefit-risk assessment based on a modified UTAUT model". Journal of Electronic Com-merce Research **14**(1), 67–84 (2011)
27. Chang, MK., Cheung, W., Cheng CH., and Yeung, Jeff HY.: Understanding ERP system adoption from the user's perspective". International Journal of Production Economics. 113(2), 928–942 (2008)
28. Yoon, K.: The Factors Affecting the Intention to Use Cloud Computing Services: Focusing on the Financial Industry. Doctoral Thesis. DanKooK University (2015)
29. Gonzalez, G.C., Sharma, P.N., Galletta, D.: Factors influencing the planned adoption of continuous monitoring technology. J. Inf. Syst. **26**(2), 53–69 (2012)
30. Lee, J.W., Kim, E.H.: Impacts of small and medium enterprises' recognition of social media on their behavioral intention and use behavior. Journal of Information Technology Service **14**(1), 195–215 (2015)

Confidential Documents Sharing Model Based on Blockchain Environment

Sung-Hwa Han

Abstract The blockchain platform has the advantage of providing shared information integrity assurance, data loss recovery, and configuration history tracing functions by applying cryptographic technology. This blockchain platform is used as a base technique for sharing information among users. However, when important information is shared through the blockchain platform, there is a disadvantage that other users can access the shared information. In the information service using the blockchain platform, it may be necessary to share defined users. Limited user's access to sharing information is allowed, but unauthorized user's access may have to be denied. To satisfy this requirement, this study proposes a model that can share confidential documents in a blockchain environment. The proposed confidential document sharing model operates on the blockchain platform and has the advantage of not using a key management server. If the proposed model is applied to the enterprise environment, it has the advantage of sharing confidential information while keeping all the advantages provided by the blockchain environment. However, since this study focused on the method on confidential file sharing, the sharing performance is inferior to other techniques, so additional research is needed.

Keywords Confidential document sharing · Blockchain platform · Distribute database · Privilege management · Limited users

1 Introduction

In an enterprise environment, various information is processed. The organization also generates information open to external and information shared by internal users. Of course they generate and use confidential information. Information is a concept and has no form. Information has a logical form only when it is saved to a database or file. The user accesses this stored information and uses the information.

S.-H. Han (✉)
Depatment of Information Security, Tongmyong University Busan, Busan, South Korea
e-mail: shhan@tu.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_10

Many organizations use encrypted documentation to share confidential information among internal users. A confidential document uses an encryption key to encrypt the document, and users who can access the document share the encryption key. Then, the user who possesses the encryption key can access the confidential document. Conversely, an unauthorized user who has not obtained an encryption key cannot access confidential documents. This is because the decryption point is when the user opens the document using the application. When the document is decrypted, the user can get the information saved in the document [1].

This document encryption function is suitable for denying unauthorized users access to confidential information saved in the document. This function can protect the confidentiality of enterprise information. However, in an enterprise environment, confidentiality is not the only requirement. Depending on the purpose and feature of information, other attributes such as integrity, availability, fault tolerance, and traceability may be required. In particular, integrity and availability of public information are emphasized [2].

The document encryption function can satisfy confidentiality, availability, and traceability, but not integrity and fault tolerance. If an authorized user has a malicious purpose, they can delete this confidential document or modify important information into other information. This is because ownership of confidential documents is assigned to authorized users. If an authorized user deletes a confidential document, the document's availability is compromised. Also, if the authorized user arbitrarily modifies the confidential document, the user who uses the modified information may get an unintended result, which may cause a problem that the enterprise goal cannot be met.

This study proposes a confidential documents sharing model based on blockchain platform as a method to protect information emphasizing such integrity and availability. The model proposed in this study guarantees confidentiality, integrity, and availability by applying the encryption function to confidential documents, and can identify and authenticate users accessing confidential documents through the encryption key management function. In addition, since confidential documents are managed on the blockchain platform, the integrity of confidential documents can be guaranteed. By sharing the user access log when necessary, it is possible to trace users who access confidential documents.

The confidential documents sharing model based on blockchain platform proposed in this study is a conceptual model. Therefore, this study includes considerations when implementing the proposed model empirically.

This study is organized as follows. Section 1 describes the background, purpose, and effect of this study. Section 2 identifies the method of processing confidential information in the enterprise environment and the security threat that occurs here, and introduces the blockchain platform. Section 3 introduces the confidential documents sharing model based on blockchain platform, which is the goal of this study, and explains considerations when implementing it empirically. Section 4 summarizes this study.

2 Related Work

2.1 Enterprise Environment for Confidential Information Process

In an enterprise environment, a lot of information is processed. Among them, confidential information is protected and monitored from the time of information creation. Confidential information is created only by the use of limited applications in the system and limited network designated by the limited user. The information created in this way is saved in the designated storage in a fixed way. Confidential information is saved in various ways depending on storage, but databases or files are mainly used [3, 4].

When the confidential information save type is file, EDRM (Enterprise Digital Right Management) system is used. The structure of EDRM is shown in Fig. 1.

Confidential documents are managed by the EDRM Controller. The EDRM Controller manages encryption keys and privileges, and saves confidential documents. The author encrypts the confidential document using a crypto tool terminal. After that, the author defines the user who can access the confidential document using the terminal, and then saves it in Rights. Rights are a list of confidential documents that each user can access. The recipient can only access documents registered in Rights. When a recipient accesses a confidential document, authentication is enforced by the EDRM controller. Then, the EDRM Controller checks the rights to see if the recipient can make a confidential document. If recipient privilege is registered, the

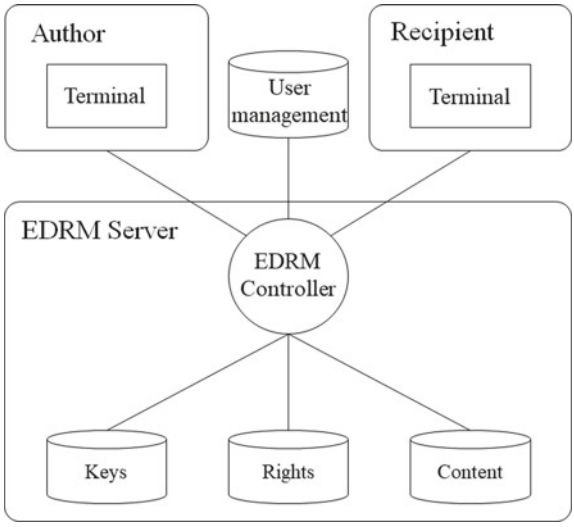


Fig. 1 Confidential information protect mechanism by DRM

EDRM Controller decrypts the confidential document using the decryption key of the confidential document that the recipient wants to access. The recipient will then be able to access the decrypted confidential document [5].

2.2 Security Threats About Confidential Document Sharing

An encrypted confidential document has the advantage of ensuring its confidentiality by accessing only limited users. However, information used in the enterprise is not always required to be confidential. Table. 1 is a representative information security requirement required in an enterprise environment [6–8].

This security requirement may require other security attributes depending on the purpose of the information and the user.

Since public information is a matter to be known to many people, confidentiality is not usually required [9]. This information was created for the purpose of being shared with other users for the public benefit. Since this information is used by everyone, everyone should share the same information. If some users or user groups use information different from other users or user groups, the stakeholder related to the user or user group may suffer damage. Therefore, since this information should not be arbitrarily changed, integrity is required [10].

In addition, users should be able to access this information using methods that are accessible at any point in time. As shown Fig. 2, when a user who needs information cannot use the information because he cannot access it, that user suffers damage. Therefore, since the user should always be able to access this information at a desired time, availability is required [6].

EDRM is a system for sharing confidential documents between internal users. If EDRM provides service all the time, availability is guaranteed. In addition, if EDRM correctly provides user authentication and encryption functions, confidentiality is also guaranteed [11]. However, integrity cannot be guaranteed.

When an internal user has access to a confidential document and has the privilege to modify the document, the user can edit the confidential document. If an internal user has a malicious purpose, other internal users may be damaged by the information the

Table 1 Iinformation security requirement required in an enterprise environment

Requirement	Description
Confidentiality	• When an unauthorized user gets the information value, the owner may suffer damage • It is necessary to protect the value of information
Integrity	• When an unauthorized user modifies or deletes information value, the user may suffer damage due to wrong information use • It is necessary to keep the configuration of information
Availability	• Information is used when the user needs it • It is necessary to maintain information status and access channels

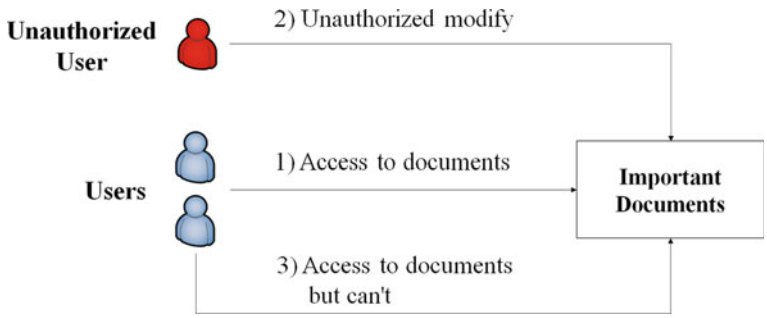


Fig. 2 Availability about confidential document

user has arbitrarily modified. This modified confidential information can be traced by the EDRM controller, but there is a problem that it cannot be restored to its previous status.

2.3 Blockchain Platform

Blockchain is a platform that can share transaction ledgers in a distributed database proposed by Satoshi Nakamoto in 2008. It is a predefined platform to implement the proposed cryptocurrency to improve the problems of electronic money [12].

This blockchain platform can theoretically share all data by nodes in the blockchain network. The shared data verifies whether the author generated it at the time of sharing. The verified data is saved in a chain method by applying the hash algorithm, and as shown in Fig. 3, it is hashed including the entire previously saved data [13].

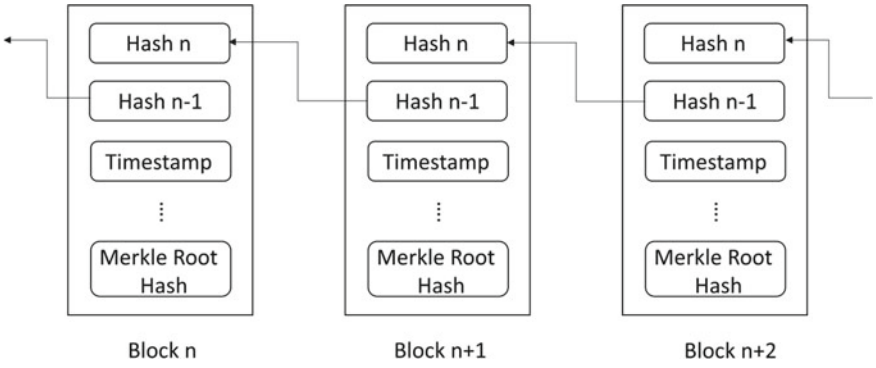


Fig. 3 Blockchain chain structure

Due to this save mechanism, it is difficult to modify the chain saved in the blockchain node. Since each blockchain node performs chain verification every time it accesses the blockchain platform, even if the chain stored in one node is tampered with, the altered chain has the advantage of being restored immediately [14].

The blockchain platform selected ledger data as sharing data. This ledger is usually text, but if you use a DApp, you can share various files such as images, audio, and video [15].

3 Confidential Document Sharing Model Based on Blockchain Platform

3.1 Security Environment Analysis and Requirements

In an enterprise environment, confidential documents may be shared among limited users, but the shared confidential documents may not be modified. A global company or an international organization shares a lot of confidential information. Also, since the time difference is applied due to different regions, the information shared for a certain period should not be changed. Not only that, even if it is modified, it should be possible to immediately restore the existing confidential document. Finally, Confidentiality, Integrity, and Availability must all be satisfied.

Although these enterprise security requirements are emphasized, EDRM, which is used by many organizations, has a problem that cannot satisfy all three requirements. EDRM can satisfy Confidentiality and Availability, and additionally, traceability. However, EDRM cannot satisfy integrity or recoverability by itself. However, all these security functions are required in an enterprise environment.

3.2 Confidential Document Sharing Model Based on Blockchain Platform

In this study, we propose a confidential document sharing model based on blockchain platform that can satisfy all of the confidentiality, integrity, and availability required in an enterprise environment.

As shown in Fig. 4, the proposed model consists of a blockchain platform, an EDRM DApp with an EDRM role, and a user terminal.

EDRM DApp is in charge of Authenticator, Privilege Manager, and Key Manager functions. The authenticator authenticates whether the blockchain node is a user who can access confidential documents. Even if it is a blockchain node, users of nodes who cannot access confidential documents are blocked from accessing them. The Privilege Manager manages the user's access authority. The Key Manager manages the encryption key required to decrypt confidential documents.

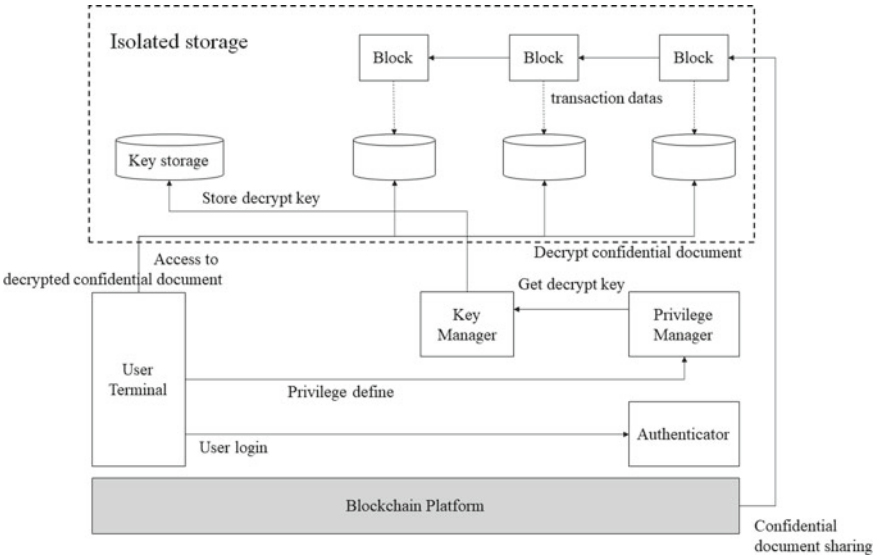


Fig. 4 Confidential document sharing model based on blockchain platform

User Terminal provides two functions. The first function is a privilege define. The document creator uses the User Terminal to define users who can access confidential documents, and configures privileges for each user. The second function is the user interface of EDRM service. By using this user terminal, user login or access decrypted confidential document.

The confidential document sharing model based on blockchain platform proposed in this study is the way EDRM function operates on the blockchain platform. Therefore, all events that occur in EDRM are shared by all blockchain nodes using blockchain transactions. In addition, the saved EDRM event is saved in a chain method, so it is safe from modification attack on the external side. If a previous version of a confidential document is needed, the previous version stored in the chain can be accessed immediately.

4 Implement

4.1 Verification Environment and Items

The confidential document sharing model proposed in this study should correctly provide the targeted security function and should not interfere with the execution of other application processes. Therefore, after empirical implementation of the confidential document sharing model proposed in this study, verify its function.

Table 2 Function verify items

Verify ID	Description
Func_01	• Verify that the distribute confidential document with other blockchain nodes • If the confidential document shareing function is provide correctly, each blockchain nodes can share other information
Func_02	• Blockchain nodes can verify received confidential document • For this function, each blockchain node should be join to blockchain network
Func_03	• Check that the verified confidential document are save as blockchain in each blokchain • If it is saved correctly, user can review that confidential document

The hardware environment of the linux system to verify the function of the confidential document sharing model proposed in this study was implemented in an i5-8500 CPU, 8Gbyte memory, and 256Gbyte SSD environment for blockchain node.

For the software environment, Redhat Enterprise Linux 8.4 was chosen for both systems. For the blockchain nodes, SSH2 and Telnet service were chosen.

The confidential document sharing model proposed in this study provides a security function that can distribute confidential documents and verify, store. Therefore, the functions verify items of the system proposed in this study are defined as shown in Table 2.

In order to implement the confidential document sharing model based on blockchain platform proposed in this study, it is necessary to structurally link the blockchain platform and EDRM. Although legacy EDRM has a CS structure, the confidential document sharing model based on blockchain platform proposed in this study has a distribute database type, so the operation environment is also different. Therefore, various matters should be considered when implementing.

4.2 Verification Process

First, a detailed analysis of the business process including the creation and distribution of confidential documents and user access is required. In a business environment, the usage of confidential documents does not occur only internally. Also, since access is not always online, various business environments must be considered.

Confidential documents of the confidential document sharing model based on blockchain platform are shared by all blockchain nodes. When the shared confidential document is saved as an application file, there is a possibility of accessing confidential information from outside the blockchain platform. Therefore, the shared confidential document should be stored in an isolated area such as a sandbox.

Key management for decrypting confident documents should also define various use cases and then define the key management process. If the key for decrypting the confidential document is exported externally, you can use this key to decrypt the

confidential document at another point in time. In addition, since the confidential document decryption key is also shared by all blockchain nodes, key storage must also be stored in an isolated area.

Documents used in an enterprise environment often contain various images. So the document size is large. Therefore, to implement the confidential document sharing model based on blockchain platform proposed in this study, it is necessary to check the storage capacity of the user device.

5 Conclusion

In an enterprise environment, there are various business processes, and these processes change frequently. The use of confidential documents in such an environment is a basic condition to prepare for business opportunity while building a safe environment. However, EDRM that provides this confidential document service can guarantee Confidentiality and Availability, but has a problem in not guaranteeing integrity.

In this study, in order to solve this problem, a confidential document sharing model based on blockchain platform was proposed and the role of each component was defined. Also, we checked the considerations to be applied when implementing the model proposed in this study.

The blockchain platform is used in various fields, and its expansion potential is very high. In particular, it shows the greatest effect in public service. We will study other application service models that can apply the many advantages provided by the blockchain platform to the business environment. Since this study focused on the method on confidential file sharing, the sharing performance is inferior to other techniques, so additional research is needed.

References

1. Krishnan, S., Neyaz, A., Shashidhar, N.: A survey of security and forensic features in popular eDiscovery software suites. Int. J. Secur. (IJS) **10**(2), 16 (2019)
2. Popescu, A., Lammich, P., Hou, P.: CoCon: A conference management system with formally verified document confidentiality. J. Autom. Reason. **65**(2), 321–356 (2021)
3. Westmacott, P., Obhi, H.: The database right: Copyright and confidential information. J. Database Mark. Cust. Strategy Manag. **9**(1), 75–78 (2001)
4. Sevak, B.: Security against side channel attack in cloud computing. Int. J. Eng. Adv. Technol. (IJEAT) **2**(2), 183–186 (2012)
5. Van Beek, M. H.: Comparison of enterprise digital rights management systems. Advice Rep., Aia Software **33** (2007)
6. Khidzir, N. Z., Mat Daud, K. A., Ismail, A. R., Ghani, A., Affendi, M. S., Ibrahim, M., Hery, A.: Information security requirement: The relationship between cybersecurity risk confidentiality, integrity and availability in digital social media. In Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016), Springer, Singapore, pp 229–237 (2018)

7. Alkhudhayr, F., Alfarraj, S., Aljameeli, B. and Elkhdiri, S.: Information security: A review of information security issues and techniques. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), IEEE, pp. 1–6 (2019)
8. Von Solms, R., Van Niekerk, J.: From information security to cyber security. Comput. Secur. **38**, 97–102 (2013)
9. Qiu, H., Qiu, M., Liu, M., Ming, Z.: Lightweight selective encryption for social data protection based on ebcot coding. IEEE Trans. Comput. Soc. Syst. **7**(1), 205–214 (2019)
10. Yuan, Y., Zhang, J., Xu, W. and Li, Z.: Identity-based public data integrity verification scheme in cloud storage system via blockchain. J. Supercomput., 1–22 (2022)
11. Joshi, K.P., Elluri, L., Nagar, A.: An integrated knowledge graph to automate cloud data compliance. IEEE Access **8**, 148541–148555 (2020)
12. Leeming, G., Cunningham, J., Ainsworth, J.: A ledger of me: personalizing healthcare using blockchain technology. Front. Med. **6**, 171 (2019)
13. Raikwar, M., Gligoroski, D., Kralevska, K.: SoK of used cryptography in blockchain. IEEE Access **7**, 148550–148575 (2019)
14. Dragomir, V. D. and Dumitru, M.: Practical solutions for circular business models in the fashion industry. Clean. Logist. Supply Chain., 100040 (2022)
15. Khatal, S., Rane, J., Patel, D., Patel, P. and Busnel, Y.: Fileshare: A blockchain and ipfs framework for secure file sharing and data provenance. In Advances in Machine Learning and Computational Intelligence, Springer, Singapore, pp. 825–833 (2021)

Distance Based Clustering in Wireless Sensor Network

Joong Ho Lee

Abstract Wireless sensor networks (WSNs) with large sensor nodes are deployed under battery constraints. Low power operation for sensor could be achieved from clustering algorithm in a field of WSNs. In a wireless sensor network, the operating lifetime of sensor nodes composed of cluster members should be maximized. Therefore, this paper proposed a hierarchical clustering algorithm that can minimize the power consumption of sensor nodes. Multi-hop routing and distance-based clustering method have proposed to improve the lifetime of sensors in the network. This article proposes a single unit distance based two-hop clustering method for clustering and analyzes the energy persistence of sensor nodes in the formed clusters. Sensor nodes clustering based on physical distances which method has many limitations in recruiting cluster members due to the limitation of the distance between cluster members and the transmission energy. The proposed cluster formation method can efficiently manage operational energy by improving the non-uniformity of cluster groups. And the simulation results of the proposed method are compared with the existing method.

Keywords Wireless sensor networks (WSNs) · Sensors node · Multi-hop routing · Distance-based clustering

1 Introduction

Micro-Electro-Mechanical System (MEMS) sensors capable of low-power operation in the field of sensor technology have been developed due to semiconductor technology capable of miniaturization. The sensor's low-power technology has played an innovative role in extending the lifespan of the sensors [1]. WSN's sensors have been developed to monitor a wide range of natural environments under limited power conditions. Sensors deployed over large areas must ensure robust operation to reliably transmit data collected in extreme natural environments [2]. Since sensors are

J. H. Lee (✉)
Depatment of AI, Yongin University, Yongin-Si, South Korea
e-mail: joongho65@yongin.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_11

deployed in a wide area, they collect data until the end of their lifespan, and it is impossible to replace batteries to extend their lifespan [3, 4]. In order to minimize power consumption in wireless sensor network technology, research on the group formation algorithm of sensor nodes is continuously being conducted. Reducing power consumption is a key factor in wireless sensor network technology [5]. As for the existing clustering technology, research on the clustering technology based on a uniform network environment has been achieved. In the real network environment, the distribution of sensor nodes may not be uniform due to the characteristics of the natural environment [6]. When sensors form a cluster in a non-uniformly distributed network, the density difference of sensor nodes occurs between cluster. The Cluster head (CH) node based on the existing absolute hop-based clustering method, collects data from member sensors node and transmits it to the destination, but the non-uniformity of the sensor nodes causes a difference in energy consumption. The CH node of a cluster with large number of sensor nodes consumes more energy than a cluster with a relatively small number of sensor nodes, which shortens the life of the cluster and consequently shortens the lifetime of the network [7]. When a CH node reaches the end of its lifespan, it cannot receive data from member sensors nor transmit it to its destination. Therefore, other sensor nodes among member nodes must be replaced with new head nodes. The shorter the replacement cycle of the cluster head node shortens the lifespan of the cluster and ultimately shortens the lifespan of the entire network [8, 9]. We should also consider the lifetime of the CH nodes in the cluster to extend the lifetime of the network. A method to increase energy efficiency, such as the clustering hierarchy (LEACH) algorithm [10], was applied to the WSN. It is a single-hop-based communication method using a single-hop clustering scheme, in which the CH node communicates directly with the sink node. However, this algorithm is less effective in saving energy for member nodes located more than 1 hop away from the sink node.

Therefore, in this study present a method to improve the cluster size for a cluster formed by non-uniformly deployed sensor nodes in WSN. In this paper, we present a clustering algorithm based on the distance between sensor nodes to enable energy-efficient clustering by forming a uniform cluster size. An algorithm for CH node selection was also presented to improve the energy efficiency of the CH node. In addition, the simulation result comparing the proposed algorithm with the existing algorithm was shown.

2 Related Review

The CH node should collect data sent from the sensor nodes in the group. If a group has n member nodes, the member node transmits data only once, but the CH node receives the data n times. Therefore, CH nodes consume more energy than member nodes. It is more energy efficient for a CH node to collect data from member node and send it to its destination (sink node) rather than each member node transmit data directly to the sink node [11]. The LEACH scheme was proposed as a way to reduce

energy consumption by electing a CH node and sending data to the sink node [12]. In order to collect data from adjacent sensor nodes and transmit it to the sink node, the LEACH scheme adopted a single-hop clustering method. LEACH periodically elects a new CH node from member nodes before the lifetime of the CH node expires. The selected CH node sends messages from members in the cluster to other nodes, and also receives messages from other nodes and forwards them to other nodes.

2.1 Radio Channel Model

To apply the data transmitter/receiver model of the wireless sensor network, E_{device} = 50 nJ/bit was applied to model the transmitter/receiver devices with $\in \in_{amp}$ 100 pJ/bit/m^2 [12]. Figure 1 shows the radio model.

In order to transmit k-bit information up to distance d in the radio model, the radio model is as follows:

$$\begin{aligned} E_{Tx}(k,d) &= E_{Tx-device}(k) + E_{Tx-amp}(k,d) \\ E_{Tx}(k,d) &= E_{device} \times k + \in_{amp} \times k \times d^2 \end{aligned} \tag{1}$$

To receive a k-bit message from the transmit node, the radio model expands as follows [12]:

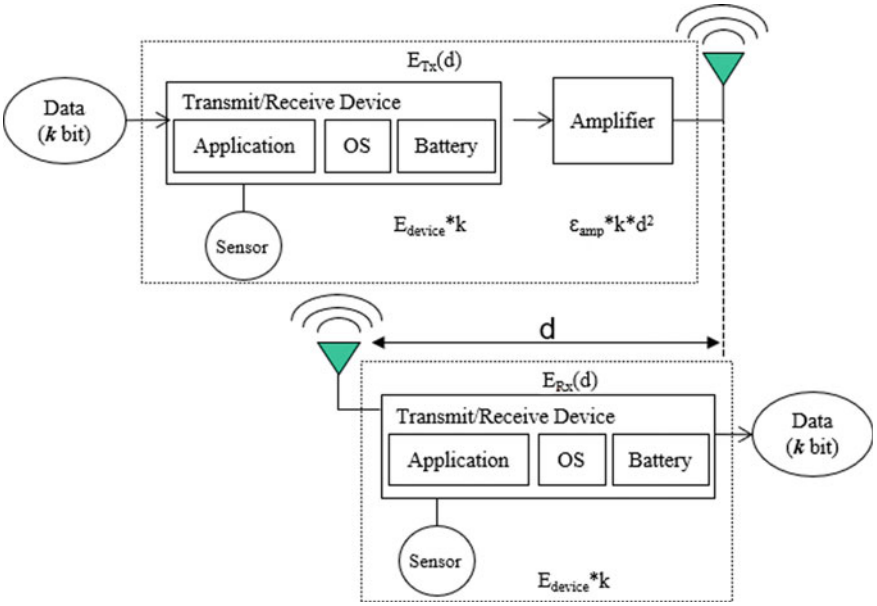


Fig. 1 Radio energy model between transmit and receive nodes

$$E_{Rx}(k,d) = E_{Rx-device}(k)$$
$$E_{Rx}(k,d) = E_{device} \times k \quad (2)$$

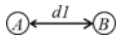
Based on the radio model, the total energy consumption depends on the routing method.

3 Cluster Formation

Clustering can form from adjacent sensor node which nodes located at adjacent distances communicate with each other to form a cluster. To complete the cluster configuration, it is necessary to select the cluster head (CH) and gate node (GW) within the cluster group member nodes. When a cluster group is initially formed, there are overlapping nodes between the groups, so group settings must be re-set so that these nodes are registered in only one cluster group.

3.1 Distance Based Cluster Algorithm

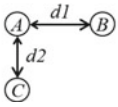
The clustering algorithm presented in this paper is based on distance. Clustering was based on the concept of forces that interact according to the distance between nodes. The strength of the interaction force defined as f_{AB} (from node A to B) according to the distance is expressed as follows by defining the physically separated distance:



$f_{AB} = 1/d1$ --- *case 1*

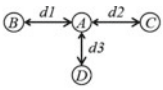
In order to represent the mutual attraction force between nodes, the distance between nodes is defined as 1 distance in d1 units. That is, the unit distance between two nodes (A–B) is d1, and the interaction force is defined in inverse proportion to the distance. Assuming three adjacent nodes, there are two types (Case 2–3) between nodes that can be formed for a unit distance in the x-axis and y-axis directions:

$f_{AB} = 1/d1, f_{BC} = 1/d2, f_{AC} = 1/(d1+d2)$ --- *case 2*



$f_{AB} = 1/d1, f_{AC} = 1/d2, f_{BC} = 1/\sqrt{d1^2 + d2^2}$ --- *case 3*

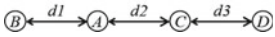
When the number of adjacent nodes is expanded to 4, the form of nodes that can be formed is the same as Case 4–7.



$f_{AB} = 1/d1$, $f_{AC} = 1/d2$, $f_{AD} = 1/d3$

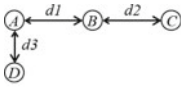
$f_{BC} = 1/(d1+d2)$, $f_{BD} = 1/\sqrt{d1^2 + d3^2}$, $f_{CD} = 1/\sqrt{d2^2 + d3^2}$

--- case 4

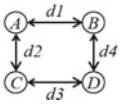


$f_{AD} = 1/(d2+d3)$, $f_{BC} = 1/(d1+d2)$, $f_{BD} = 1/(d1+d2+d3)$

--- case 5



$f_{BD} = 1/(\sqrt{d1^2 + d3^2})$, $f_{CD} = 1/(\sqrt{d3^2 + (d1 + d2)^2}$ --- case 6



$f_{AD} = 1/(\sqrt{d1^2 + d4^2}) = 1/(\sqrt{d2^2 + d3^2})$

$f_{BC} = 1/(\sqrt{d1^2 + d2^2}) = 1/(\sqrt{d3^2 + d4^2})$ --- case 7

In the same way, by expanding the number of neighboring nodes, it is possible to consider variously distributed neighboring nodes.

3.2 Node-to-Node Cohesion

In the previous section, a simplified model of mutual tension between nodes based on distance was shown. To form a cluster group, cohesion between nodes is considered as a model. For example, it shows that if a mutual force is applied between nodes from the initial state of case 1, cohesion inter nodes occurs and can be changed to the next state. Figures 2 and 3 shows the state of realignment of nodes considering cohesion for each case. Figure 2 shows the configuration in which A and C gather around node B. Figure 3 shows the cohesion form when B and D are located at right angles with respect to node A and node C is adjacent to node B.

The process of cohesion considering the mutual influence between adjacent nodes is an effective concept to form an appropriate cluster. Figure 4 shows the example of the process of group formation for randomly deployed sensor nodes. The neighboring

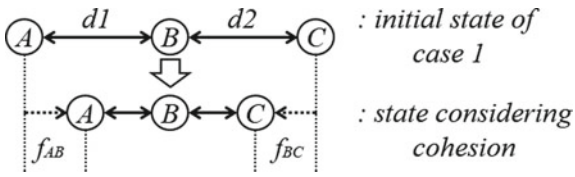


Fig. 2 Rearrangement state for the case 1 from initial state

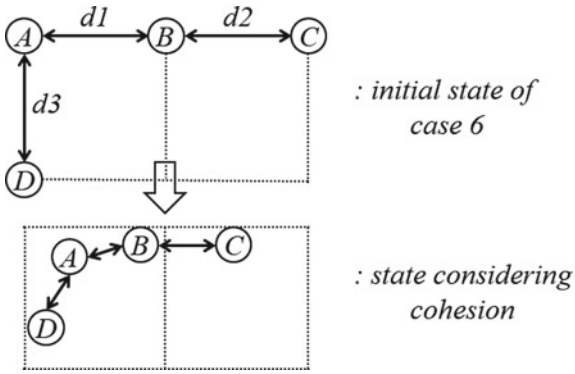


Fig. 3 Rearrangement state for the case 6 from initial state

nodes have diagrammed the process of gathering nodes located close to each other which is unit 1 distance. In this paper, to simplify the distance between adjacent sensor nodes, the unit distance is set to 1. The distance between the closest nodes is set as the unit distance 1. That is, the distance between adjacent nodes horizontally and vertically is 1, and the diagonally located nodes are also included in the unit distance 1. Figure 4a shows an example of randomly distributed adjacent sensor nodes, and Fig. 4b shows the results of aggregation into appropriate groups to form a cluster. And Fig. 4c shows the results of cluster groups. For a node at coordinates (4, 4), adjacent cells affect this node with similar strength, but eventually become incorporated into a strongly pulling group. Consequently, three groups are formed and located at coordinates (4, 4) node is registered as members of group B. In this way, cluster can be formed by applying an interaction force between a node and its neighbors.

Figure 5 shows an example in which about 1000 sensors are randomly distributed in a 100 × 100 unit area and then a cluster is formed between adjacent nodes. Figure 5 shows simulation results of the proposed clustering algorithm. The cluster formation consists of two steps. First, to form a cluster from randomly scattered sensor nodes, each node searches for a node located at a distance of unit 1 (1d) between each node.

If they find unit 1 distance node, they communicate with each other and note the id number of each node for node identification. Second, each node searches for adjacent nodes that extend to unit 2 distance (2d). In this simulation, the distance between nodes affected by reciprocal attraction is limited to 2d.

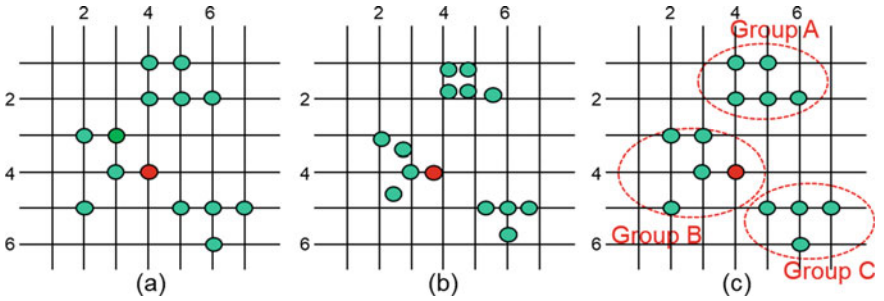


Fig. 4 Example of group formation

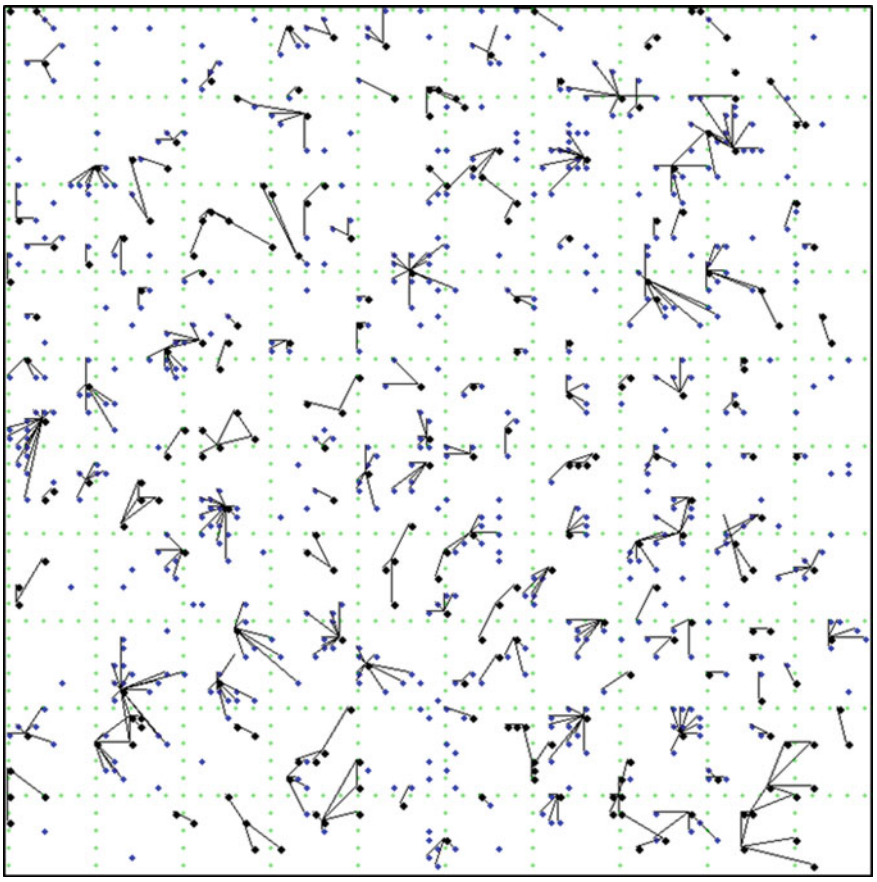


Fig. 5 Sensor node clustering simulation for a 100 × 100 Unit Area

3.3 Clustering

In order to form a cluster group between adjacent nodes in the distance-based model of adjacent nodes as in the cases described previous, the following process is performed.

p1. Every node sends a broadcast message containing latitude and longitude information to neighboring nodes.

p2. Each sensor node receives a broadcast message from the nearest node. The received message contains latitude and longitude information. It registers first for the earliest received message and calculates the distance.

p3. Sends a group member request message to the sensor node of the received message.

p4. Receive acknowledgment message from node that requested group member by p3.

p5. When node receive an acknowledgment message, node register as a member of itself.

Group formation for all the sensor nodes is completed by performing the above processes p1–p5.

3.4 Cluster Uniformation

Sensors are randomly deployed over a large area under hostile weather conditions. Due to the random distribution, the sensors are not evenly distributed. So, some clusters have a high density distribution and some clusters have a low density distribution. The more sensors are located adjacent to each other, the higher the probability that the size of the cluster will increase.

The sensor node communicates with the CH (Cluster Head) node to send data to the CH node in the same cluster, and the CH node sends the collected data from the member sensor node to the sink node. If the size of the cluster is large, the amount of data to be collected from the member nodes in the cluster increases, and consequently, the CH node consumes more batteries, and the lifespan is shortened. As the battery consumption of the CH node becomes faster, the lifespan is shortened, and the CH node election from member nodes in the cluster may become frequent. Accordingly, the life span of all member nodes in the cluster can be shortened quickly. Therefore, if the size of a cluster is larger than that of other clusters, the size of the cluster can be made uniform by dividing the cluster. When the size of the cluster is reduced, the amount of data to be collected by the CH node from the member nodes is reduced, thereby reducing battery consumption. In the distance-based clustering process, the process of dividing the cluster when the size of the cluster is large is shown in the Fig. 6. Figure 6a shows an example of a group in which adjacent member nodes are composed of unit distance 1. Group 1 composed of excessively dense members

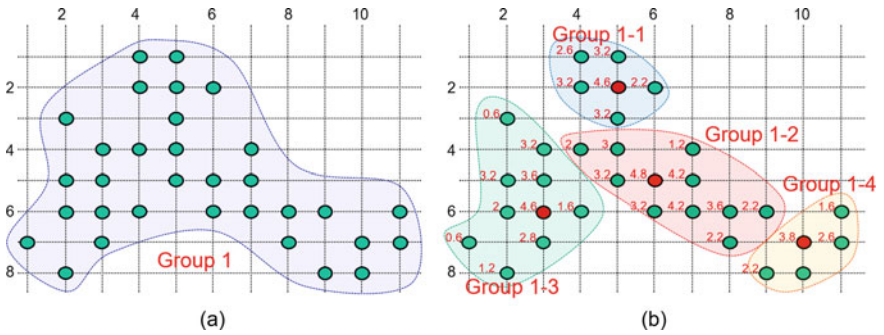


Fig. 6 Group division for a group with a high density of members within the group

can be divided into 4 groups (Group 1–1 to Group 1–4) as shown in Fig. 6b. In Fig. 6b, the interaction force according to the distance between each sensor in the same group is expressed numerically. The sensor located at the coordinates (6, 3) has four vertically and horizontally adjacent sensors and one diagonally adjacent sensor, so the magnitude of the interaction force is calculated as 4.6. Unit distance 1 included horizontally and vertically adjacent sensors and diagonally adjacent sensors. The horizontal and vertical interaction force was calculated as 1 and the diagonal interaction force was calculated as 0.6.

Figure 7 shows an example of many groups form in which about 1000 sensors are randomly distributed in a 100 × 100 unit area and then a cluster is formed between adjacent nodes. Figure 7a shows simulation results of the un-uniformed group. Figure 7b shows the results with the groups being properly uniformized. In the simulation result, the black dot indicates the sensor node, the red dot indicates the CH node which is elected from the member nodes within the same group, and red solid lines indicate that each member node in the same group is connected to CH.

3.5 Cluster Head Election—Scheme of Finding Center of Gravity Node

After cluster formation, each cluster group has to elect a CH node. Each node has connected to the nodes among the members of the cluster group. From this connectivity, the node having maximum connectivity among the members can be found. This node is selected to be a CH node. The process of selecting the CH node from the formed cluster is shown below.

p6. For all nodes within the group, each node calculate the interconnection strength between the member nodes of the group.

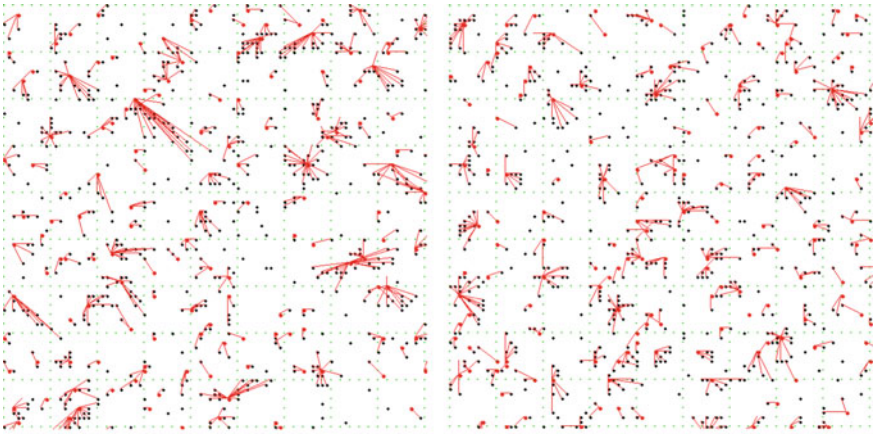


Fig. 7 Simulation results for group division for a group with a high density of members within the group

p7. Find the node corresponding to the center of gravity in the group. The node with the greatest connection strength among member nodes corresponds to the center of gravity in the group.

p8. The node with the weakest connection strength is removed first.

p9. Repeat p6 until there are no adjacent nodes.

p10. The last remaining node is elected as a CH node and broadcasts the CH node to its members.

p11. All member nodes register the CH.

The selection of CH nodes is completed by repeating the process from p6 to p11. Figure 8 shows the process of selecting a CH node from its members, assuming a specific cluster. The process of excluding cluster members one by one in each step is shown for a specific cluster group as shown inf the Fig. 8. Finally, the node at coordinates (3, 4) is elected as the CH node.

3.6 Energy Consumption Modeling of CH Node

If the CH (Cluster Head) has a shorter lifetime than other member nodes during the data collection and transfer operation after cluster group completion, the cluster head must be re-elected among other member nodes in the cluster group. Initially, assume that all of the node has an energy level 1 and the energy level is reduced by 0.1 for data collection and transmission. The member node should only transmit data to the cluster head node, and the cluster head should transmit the collected data

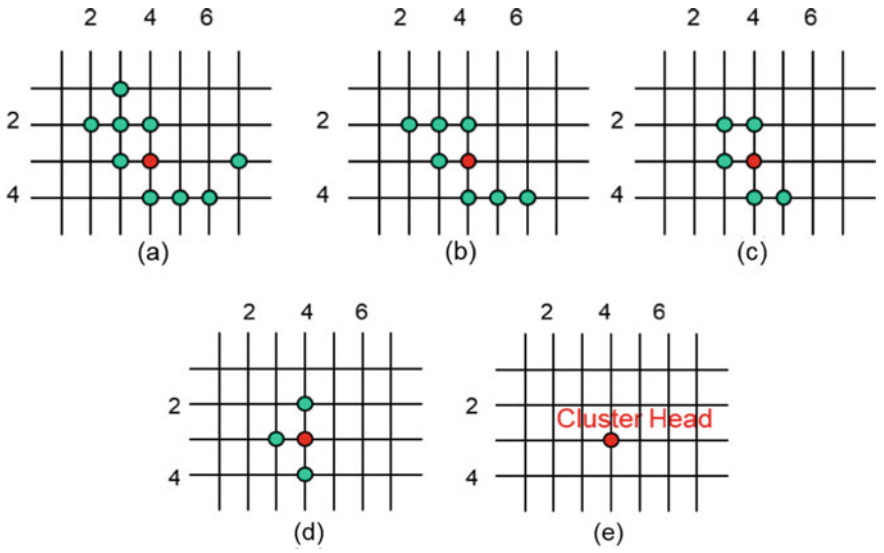


Fig. 8 Example of CH node election

from the sensor node to the base station. CH node consumes much more energy than sensor node. Figure 9 shows the energy consumption results for each CH nodes in the group. To predict the battery consumption of the CH node in the group, this paper makes several assumptions. It is simply assumed that the CH node consumes 0.01 J/time step of energy when collecting data from each sensor node once. When election a CH node, additional energy consumption is required because selection signals must be received from all member nodes in the group. Initially, the lifetime value of all nodes is set to 1, and at every time step, the CH node consumes the energy required to collect data from all member nodes, so the amount of energy is subtracted. Based on the energy consumed by each CH node for the group shown in Fig. 9, the results are shown until the remaining battery power reaches 0. Group 1 consists of 26 sensor nodes, and it is assumed that the CH node consumes 0.26 J of energy when collecting data from these sensor nodes at each time step. In conclusion, the battery of the CH node was exhausted within 4 time steps of Group 1. In the case of group 1–4, there were 5 member nodes and it was performed for 20 time steps. 5 times longer operation time than Group 1 is guaranteed. Figure 9 shows the result of simply calculating the amount of battery consumed by the CH node when the CH node receives data from the member node in the group for each time step.

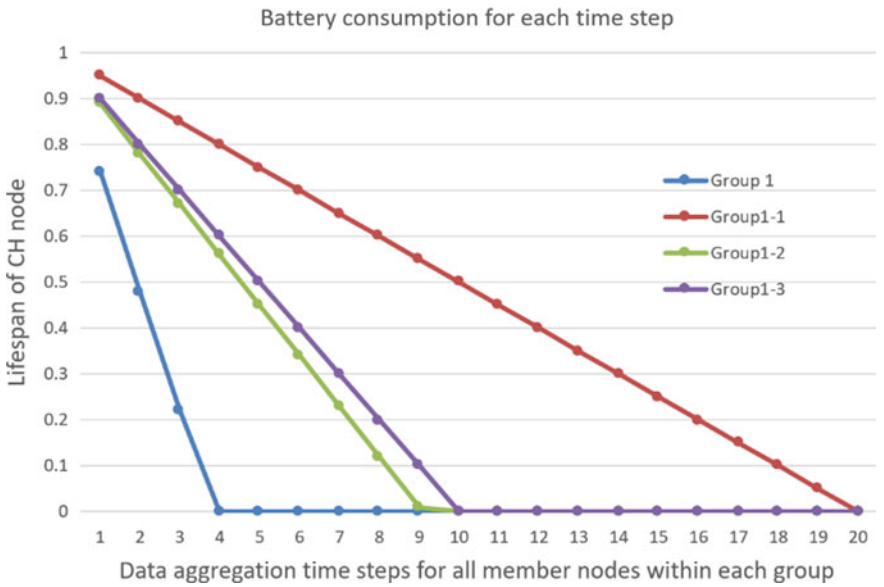


Fig. 9 Lifespan of CH node in the group

3.7 Cluster Uniformity

Consisting of a uniform cluster is an essential in extending the network lifespan. The overcrowding of the cluster group due to the uneven distribution of sensor nodes is the cause of the cluster non-uniformity. In this study, it was possible to uniformize the size of the clusters through the second clustering process from the first stage of cluster formation. By reconfiguring the CH node for a group with a higher density than the adjacent cluster group, group can be divided into two or more groups.

4 Simulation Results

The size of each group member node for each generated cluster group was confirmed through simulation. Figure 10 shows the distribution of the size of the sensor nodes configured in the created cluster group. This is the result when it is repeatedly created 3 times, and most of the member nodes of each cluster consist of 1 to 10.

Figure 11 shows that about 1000 nodes are created and the results of 100 iterations results are accumulated. At this time, the node distribution area was set to 100×100 unit area. Comparing the results without cluster uniformity and after equalization, it can be seen that there is an effect of improving the cluster size by about 20–30%.

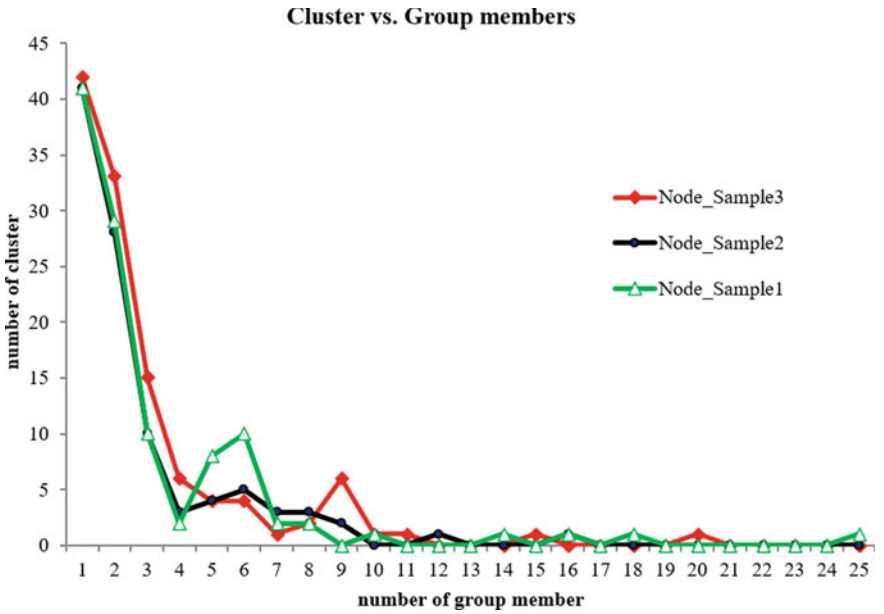


Fig. 10 Simulation Results for Number of Groups with the Same Number of Members

5 Conclusion

This paper describes a clustering algorithm based on distance and shows the algorithm of distance-based CH node election. And shows the simulation results of cluster formation in Fig. 5. Proposed algorithm is based on the concept of reciprocal attraction between nodes. We started with the assumption that there is a pulling force (reciprocal attraction) between nodes. Based on this assumption this paper presented clustering and cluster head (CH) election algorithm. By finding the node corresponding to the center of gravity of the group, it can be selected as a CH node. This procedure is shown in the p6–p11. In the clustering stage, this paper proposed resetting the CH node for overcrowded sensor nodes to improve the uniformity for the ununiformed groups. Divide of the group for the overcrowded member can improve the battery consumption of the CH node by solving the overcrowding of member nodes in the group. Figure 5 shows the clustering simulation results for 1000 random nodes. The proposed clustering method makes it possible to distribute members evenly by applying distance between sensors. Figure 7 shows the simulation results when there is an overcrowded groups and the overcrowded group is divided into small groups. In conclusion, it was confirmed that the overcrowded group could be improved by about 30%. As shown in Fig. 10, we obtained improved results for the uniformity of the clusters. The degree of uniformity can be improved by

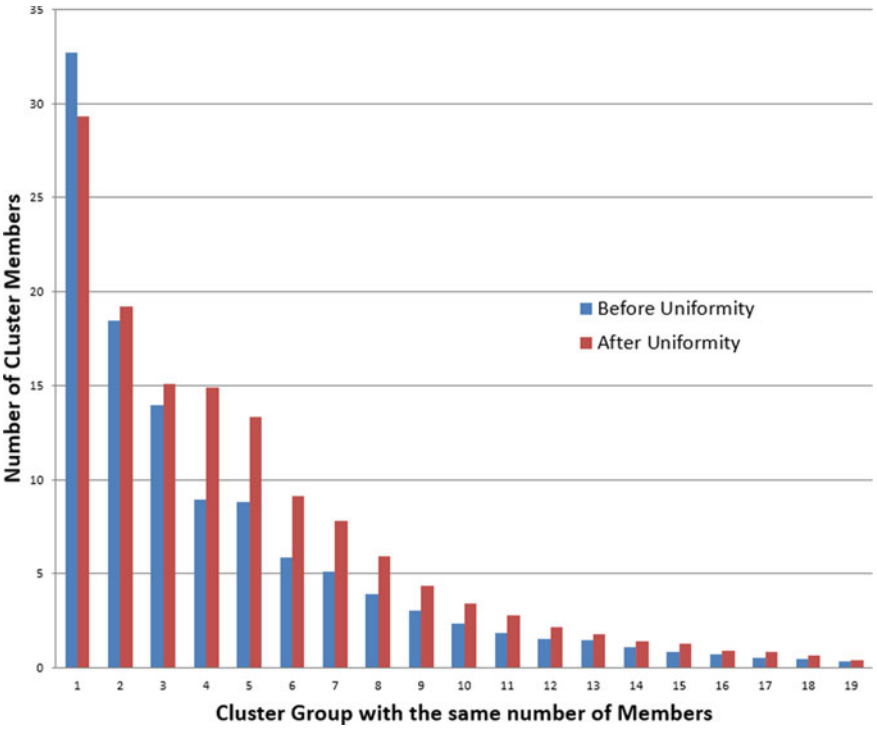


Fig. 11 Simulation Results for Cluster Groups Uniformity with the Same Number of Members

repeating the clustering step. In the case of ungrouped sensor, it reduced up to ~30%. The proposed clustering algorithm is effective for clustering of WSNs because it is simple to implement and is expected to be effective for low power implementation.

References

1. Md. Zair Hussain, M. P., Singh, R. K.: Singh Analysis of Lifetime of Wireless Sensor Network. Int. J. Adv. Sci. Technol. **53**, 117–126, (2013)
2. Hill Jason Hill, J., Szewczyk, R., Woo, A., Hollar, S., Culler, D., Pister, K.: System architecture directions for networked sensors. In: Proc. Ninth Int'l Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS '00), pp. 93–104 (2000)
3. Mukhopadhyay, S., Panigrahi, D., Dey, S.: Model based error correction for wireless sensor networks. Sens. Ad Hoc Commun. Netw. (2004)
4. Jianzhong, L., Hong, G.: Survey on sensor network research. J. Comput. Res. Dev. **45**(1), 1–15 (2008)
5. Liu, C., Wu, K., Pei, J.: An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation. IEEE Trans. Parallel Distrib. Syst. **18**, 1010–1023 (2007)

6. Wong, S., Lim, J., Rao, S., Seah, W.: Multihop localization with density and path length awareness in non-uniform wireless sensor networks. In: Proc. Of the International Symposium on Parallel Architectures Algorithm, and Networks, vol. 4, pp. 2551–2555 (2005)
7. Kim, E., Kim, D., Park, J.: Min-distance hop count based multi-hop clustering in non-uniform wireless sensor networks. Int. J. Contents **8**(2), 13–18 (2012)
8. Heizelman, W., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wireless Commun. **1**(4), 660–670 (2002)
9. Saleh, S., Al-Awamry, A., Mahmoud, F. M.: Energy-efficient communication protocol for wireless sensor networks. Int. J. Eng. Res. Technol. **4** (2015)
10. Younis, O., Fahmy, S.: HEED: a hybrid energy-efficient distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**(4) (2004)
11. Jeon, H., Park, K., Hwang, D.-J., Choo, H.: Sink-oriented dynamic location service protocol for mobile sinks with an energy efficient grid-based approach. MDPI Sensors **90**, 1433–1453 (2009)
12. Heinzelman, W. R., Chandrakasan, A. P., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proc. of the Hawaii International Conference on System Sciences, pp. 3005–3014 (2000)

Study on OSINT-Based Security Control Monitoring Utilization Plan

Dain Lee and Hoo-Ki Lee

Abstract Recently, as cybercrime and security threats that abuse the anonymity of the dark web and various social media continue to increase, the need for data analysis through OSINT has emerged. However, the reality is that public institutions and private companies are far short of response strategies compared to the increasing number of cases of infringement and public information that collect vast amounts of data. Today, with the development of the Internet, OSINT collects information from various public sources such as deep web and dark web as well as surface web such as social media, online community, and Internet news, so it must be able to monitor, collect and prevent public information using OSINT to effectively prevent leakage and infringement. In this study, the definition and concept of OSINT, and utilization plans in security monitor including characteristics are studied and presented, and OSINT information collection tools are identified and analyzed. Through this, OSINT-based security monitor utilization plans and considerations are presented.

Keywords OSINT · Social · Development · Security monitor

1 Introduction

Among the information media developed by mankind, the Internet has the largest variety and amount of information, delivers it throughout the world at the fastest pace, and has become a convenient and inexpensive information access medium shared by the largest number of people. Ultimately, all the public information that humanity needs is expected to exist on the Internet. As much of the information that had to rely on collectors in the past can now be collected in real time through the Internet, the act of abusing public information is increasing day by day [1].

D. Lee
Department of T&D Security, Daejeon-Si, South Korea

H.-K. Lee (✉)
Department of Cyber Security Engineering, Konyang University Nonsan-Si, Nonsan-Si, South Korea
e-mail: hk0038@konyang.ac.kr

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_12

Recently, the need for data analysis through OSINT has emerged as cybercrime and security threats that exploit the anonymity of the dark web and various social media continue to increase. About 533 million personal information leaked from social network services in 2021 was released free of charge on the Dark Web Forum, of which 120,000 were also included in Korean information. In March 2021, Doppel-Paymer, a cyber attack organization, carried out malicious code attacks on global automobile manufacturers, which leaked internal information such as email backup files. They attempted to make financial demands by releasing the leaked information little by little on the dark web. Information disclosed in such public source information can be feared to cause secondary damage by exploiting it. Currently, it is possible to easily penetrate servers, databases, and IoT devices with simple filtering searches using public tools. However, the reality is that the response strategies of public institutions and private companies are far insufficient compared to the increase in public information and infringement cases that collect vast amounts of data. Institutions dealing with important information need to collect and analyze information from various sources including the dark web and the deep web. OSINT is the easiest way to get information because it uses public information accessible to anyone [2], but finding accurate and significant content from the vast amount of information spread on the real web, and tracking the identity of criminals behind anonymity requires a lot of time and know-how. In particular, it is virtually impossible for the general public to monitor spaces that can be used as open sources like the dark web, but are difficult to access, unlike the surface web, and collect necessary information. Since OSINT collects information from various public sources such as deep web and dark web as well as surface webs such as social media, online communities, and Internet news, OSINT should be used to monitor, collect, and prevent public information.

This study researches and presents ways to use OSINT in security monitor, including the definition, concept, and characteristics of OSINT, and identifies and analyzes current OSINT tools. Through this, OSINT-based security monitor utilization plans and considerations are presented.

2 OSINT Definitions and Concepts

2.1 Definition of OSINT

OSINT stands for Open Source Intelligence, which is a combination of Open Source, which means a public source, and Intelligence, which means information collected for purpose through filtering, analysis, and processing, and refers to information obtained from public sources. OSINT is also called open source intelligence or public information, disclosed information, and open source information.

OSINT began to be used by countries and private intelligence organizations to identify information related to terrorism and criminal activities, and began to collect open-source information through traditional media such as newspapers, radio, and

TV. In addition, it is used for the purpose of establishing an effective threat strategy to protect against cyber attacks. Blogs, e-mails, IP addresses, SNS, and dark web are being used for the purpose of collecting specific information.

2.2 Concept of OSINT

OSINT is a collection of data in a public and legal manner, It includes information obtained through public observation on foreign political, economic, and military activities, and information obtained from radio and television.

The concept of information is often used without distinction between "Information" and "Intelligence," but it is actually divided into different meanings. Information obtained through the fact-checking stage is defined as "Intelligence" for public information, and only information that has been analyzed and evaluated is referred to as "Intelligence". Data collected using public information at the stage before analysis and evaluation are defined as "information". OSINT is a method of information collection, which refers to information collection activities from the collection stage to the information being useful through analysis and evaluation [3].

2.3 Features of OSINT

The types of information collection can be classified into HUMINT, TECGINT, and OSINT, and the characteristic of OSINT is to produce useful information by collecting and analyzing information from data from public sources such as Table 1.

The type of OSINT can be classified into two perspectives: Offensive and Defensive in terms of security. Offensive refers to directly collecting information by searching for keywords, etc., and Defensive refers to checking the collected information [4] (Fig. 1).

OSINT has the advantage of being able to acquire information relatively quickly, ensuring the speed of data access. In addition, it is characterized by securing more data than secret intelligence data. It can be obtained at a low cost, and anyone can

Table 1 Comparison of input and output variables in previous research

Sortation	content
HUMINT	It is an information collection activity technique that collects information using human information and interpersonal contact, and refers to the information itself collected by human assets
TECGINT	Technology information activities are activities that collect information using advanced technologies, not human information sources
OSINT	Information gathering techniques that produce useful information by collecting and analyzing information from public sources available to anyone without restriction

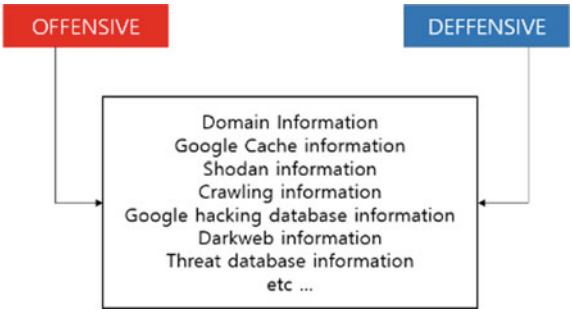


Fig. 1 Comparison of input and output variables in previous research

easily access and use it conveniently. Of course, OSINT's characteristics do not only have advantages. Excessive amounts of data, organizational and cultural biases in the information community, security issues, and technical constraints should be con-sidered before effective use or use [5].

2.4 Application Field

Information processed by OSINT's intelligence activities is used in many areas to identify hidden facts related to national security threats, terrorism, and criminal activities. In addition, it is widely used in fields such as Internet betting and Crime Investigation. Examples include three applications: cybercriminal and organized crime, social opinion and sentiment analysis, and cybersecurity and cyber defence, as shown in Fig. 2 [6].

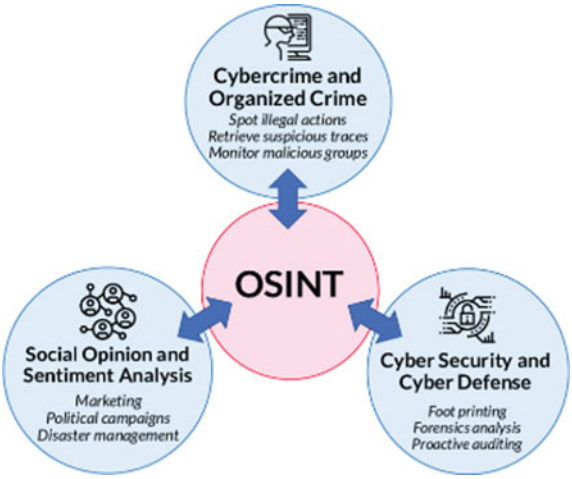


Fig. 2 OSINT field of application

3 OSINT Processes and Tools

3.1 OSINT Process

INFOSEC proposed the basic structure of the OSINT process as shown in Fig. 3 to efficiently collect information using OSINT [7].

In the basic structure of the OSINT process proposed by INFOSEC, relevant data are collected from the identified sources through the step of identifying the source, filtered, and final results are derived.

However, the OSINT process is not limited to one, but is gradually modified and used depending on the situation. If additional information is needed in the analysis stage analysis of the basic structure of (Fig. 3), the OSINT process of the new structure can be constructed by repeating the information collection, processing, and analysis steps to find and refine the association between the information (Fig. 4) [8].

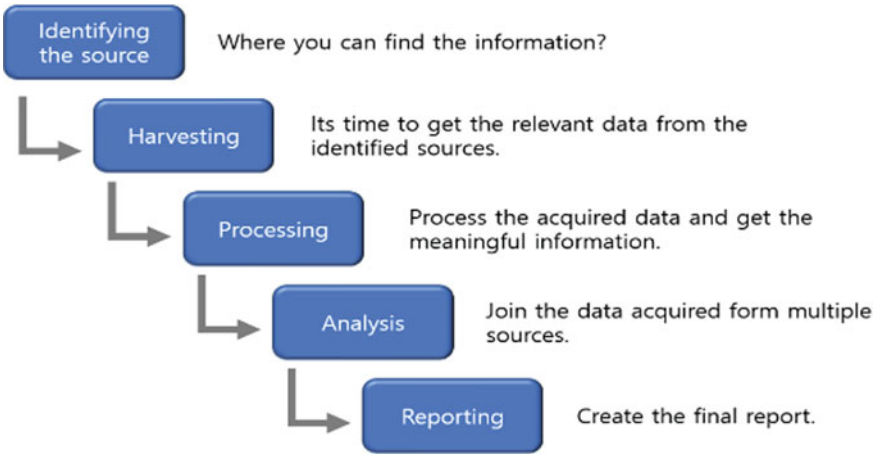


Fig. 3 Basic structure of OSINT process

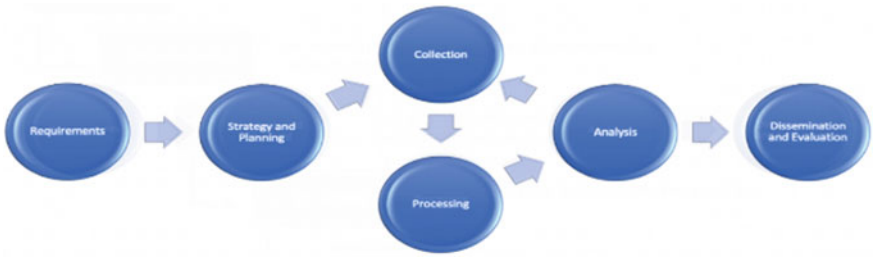


Fig. 4 Application Structure of OSINT Process

3.2 OSINT Tools

If you need to collect the necessary information on the web, it takes a lot of time to analyze it on search sites and professional sites until you get the correct results. However, when using the OSINT tool, it is desirable to actively use the OSINT tool because it is provided in various forms such as text format, file, and image and can collect and process information within seconds. Knowledge Nile selected 'Top 14 OSINT tools as of 2021' as shown in Table 2.

Censys

Censys is an OSINT search tool site that can collect information about hosts and networks, helping to continually discover unknown assets and solve Internet problems facing risks, and performs weekly scans worldwide using a Zmap scan method.

Censys performs ping operations through ZMap and ZGarb, which query information on numerous systems connected to the external Internet, to recognize which type of device responded and to identify details such as configuration and encryption. As shown in Fig. 5, Censys generates and manages data by performing scan tasks and important field extraction, structured data generation, Zdb central management, and recording procedures.

Table 2 Top 14 OSINT tools as of 2021

No.	Name	No.	Name
1	OSINT Framework	8	Metagoofil
2	theHarvester	9	Aircrack-ng
3	SHODAN	10	Censys
4	Searchcode	11	Google Dorks
5	Nmap	12	ZoomEye
6	SpiderFoot	13	Maltego
7	ExifTool	14	BuiltWith

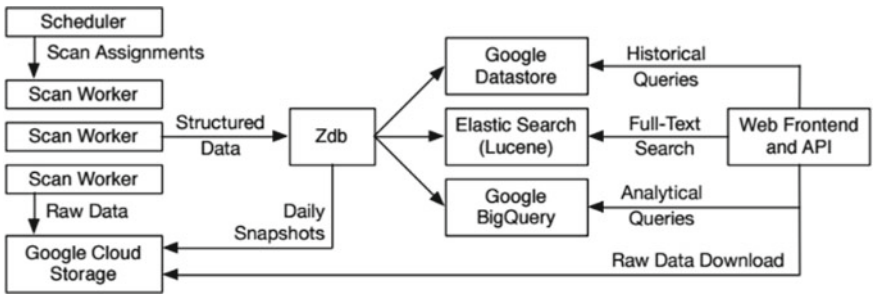


Fig. 5 Censys system architecture

Table 3 Shodan filter list

Filter	Explanation
City	Limit search results to within a given city
Country	Limit search results to within a given country
geo	Search with specific latitude/longitude information
Os	Check search results for a specific OS
Port	Check search results for a specific port
Hostname	Provides results that match a given hostname in the search results
Net	Check search results only for specific classes
Before/after	View search results before and after a specific date

Shodan

After scanning the port, Shodan is a tool that collects metadata from banner information and collects and displays information about devices connected to the external Internet, and can identify sensitive information such as IP addresses and network connectivity as well as equipment location information. Shodan's original purpose was to make it easier for security professionals to find vulnerabilities, but it is now also used to collect information maliciously. Like a general search engine, Shodan can obtain data through search, but it has limitations in viewing results depending on whether or not it is a paid/free member. The search method adopts a search method by filtering, and supports various filters such as country, city, os, hostname, and port, as shown in Table 3.

Unlike search engines that generate indexes on web content, Shodan operates in the form of collecting information through banner grabbing through open ports on servers or equipment that use the Internet, creating indexes, and returning them to clients. The returned data provides necessary information such as equipment information and services to the user through the Shodan Web UI.

SpiderFoot

SpiderFoot is one of the top OSINT tools designed to automate OSINT information collection. In addition to automatically collecting data from more than 100 sources to manipulate key scans and support information retrieval, it also includes modules equipped with data investigation functions such as Fig. 6.

Information collected through SpiderFoot includes IP addresses, domain/subdomain names, hostname.network subnet (CIDR), ASN, email address, phone number, and user name, and integrates with almost any available data source to focus on data analysis.

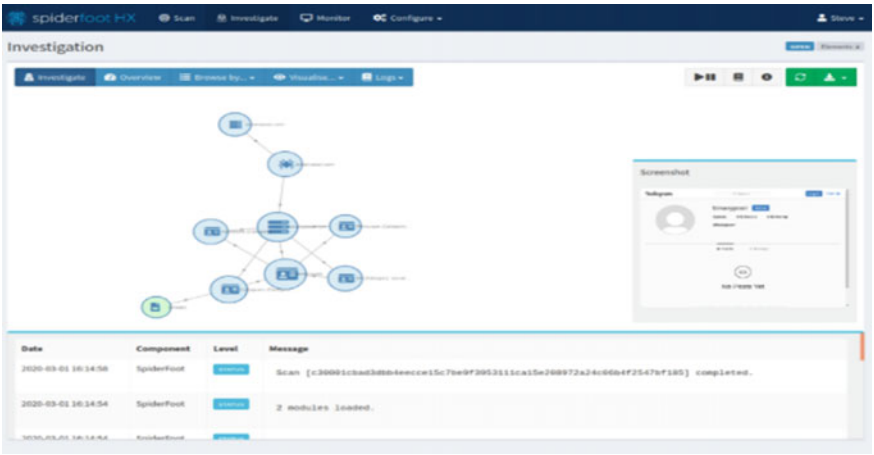


Fig. 6 SpiderFoot data investigation capability

4 OSINT Security Monitoring Utilization Plan

The use of OSINT in terms of security monitoring can include various activities, and many OSINT tools have been developed to support this activity. Most tools support more than one function, but since they are often specialized in one function, OSINT tools and target selection methods should be reviewed to facilitate the use of security monitoring on the OSINT basis [9].

4.1 Collecting and Taking Action on Internal Asset Information

The security monitoring can safely protect internal assets by monitoring and checking the leakage status of internal asset information and performing measures using the OSINT information collection tool. When collecting information based on IP, the risk of infringement should be reduced by checking open port information, server and version information exposure, service exposure, and presence of vulnerabilities.

The following Fig. 7 is an OSINT-based internal asset information leakage collection and action process. After selecting the inspection target and inspection scope based on internal assets, OSINT tools to be used for information collection are selected and carried out. Using the selected inspection method and OSINT tool, the security monitoring center collects information and utilizes the calculated results to implement security measures such as removing vulnerabilities, blocking ports, and removing services.

It is also a good idea to check the exposure status of internal domains such as administrator pages and intranets using search engine information collection tools

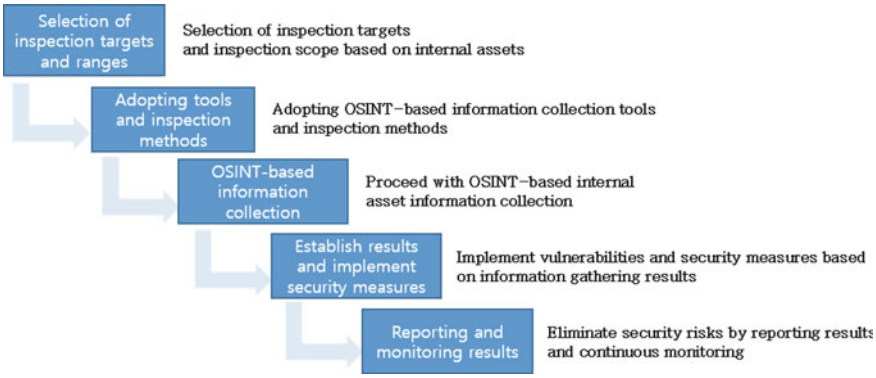


Fig. 7 OSINT-based internal asset information leakage collection and action process

such as Google Dorks. It is very important to collect and monitor OSINT-based internal asset information at all times to eliminate possible security risks in advance.

4.2 Gather and Prevent Published Threat Information

In security monitoring, the OSINT information collection tool can be used to collect disclosed threat information and vulnerabilities and block security risks in advance, thereby protecting internal assets.

As shown in Table 4, the disclosed threat information includes Malware Information, new vulnerability information, and malicious domains, and in addition, numerous information can be obtained using the OSINT information collection tool.

The following Fig. 8 is an OSINT-based open threat information collection and prevention process. It is very important to collect threat information using the OSINT information collection tool, and then primarily to identify damage from threat information collected through event information and log analysis. Depending on the occurrence of damage, accident investigation and recovery work may be required, and finally, it is important to block security threats collected through the creation of new patterns and registration of blocking policies.

Table 4 Types of threat information disclosed

No.	kind
1	Malware Information
2	New Vulnerability Information
3	Phishing Site Information
4	Maliciousness IP, URL
5	Malicious mail Information

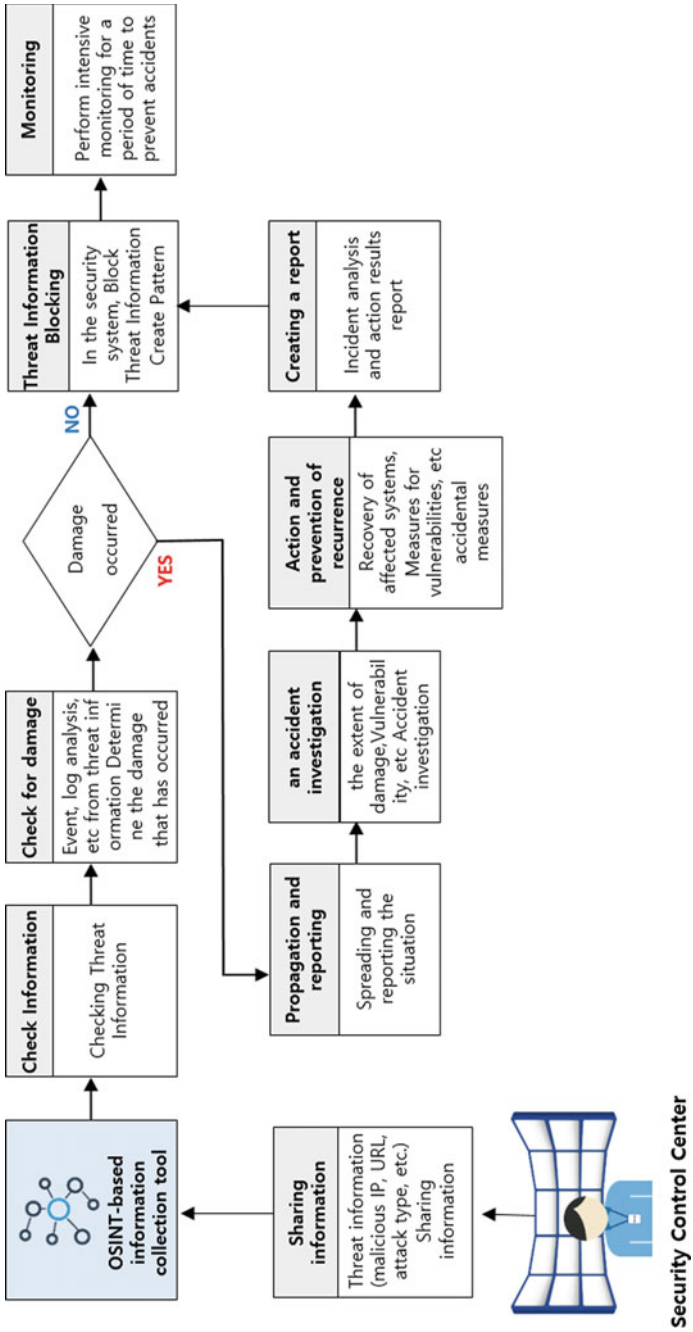


Fig. 8 OSINT-based published threat information collection and prevention processes

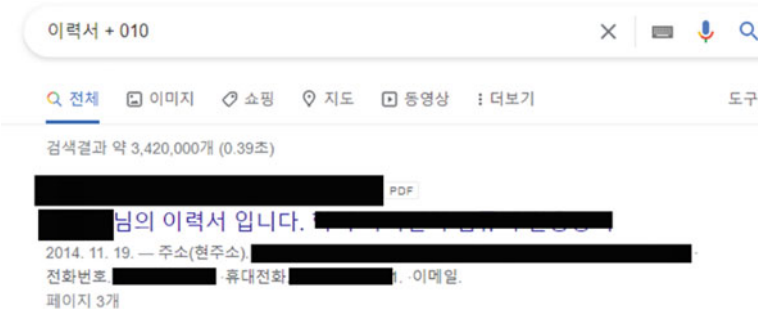


Fig. 9 Examples of personal information disclosure

4.3 Inspection of Exposure of Important Information

With the development of search engines, information can be searched quickly, but as there are vast amounts of data, damage such as using search engines for personal information hacking purposes from outside or exposing important data on search engines is continuing [10]. As shown in Fig. 9, it is possible to browse the file of a document containing personal information even with a simple keyword search.

Therefore, security monitoring should check and take action on the exposure of important information such as personal information and documents using search engine information collection tools such as Google Dorks among OSINT information collection tools to prevent leakage of internal personal information and important data. When conducting an information exposure check, it should be possible to proceed using 'filtering' to distinguish vast amounts of data. After entering the internal asset domain or IP as follows, the keyword can be combined to proceed with the check.

OSINT-based critical information exposure checks are very important to perform monitoring at all times to eliminate possible security risks in advance.

5 Conclusions

Since OSINT collects information from various public sources such as deep web and dark web, it is necessary to monitor, collect, and prevent public information using OSINT information collection tools in terms of security monitoring in order to effectively prevent leakage and infringement. However, in order to accommodate this, it is necessary to establish an information collection and response strategy that meets the organization's goals and vision, and to establish a process for utilizing OSINT-based security monitoring.

In this study, a method that can be used in security monitoring was studied for this content. An overview of the definition, concept, and utilization plan of OSINT

was presented, while the characteristics of various OSINT information collection tools were presented. In addition, as OSINT-based security monitoring utilization measures, internal asset information collection and action, public threat information collection and prevention, and important information exposure check were classified and studied.

Based on this study, it is hoped that by realizing OSINT-based security system utilization measures, invisible threats can be prevented and a more strengthened security system can be established.

Acknowledgements This paper was supported by the Konyang University Research Fund in 2022.

References

1. Byungchul, C.: A system for national intelligence activity based on all kinds of OSINT (Open Source INTelligence on the Internet) (2003)
2. Kyuyong, S., Jincheol, Y., Changhee, H., Kyoung, M. K., Sungrok, K., Minam, M., Jongwan, L.: A study on building a cyber attack database using open source intelligence (OSINT) (2019)
3. Wanhee, L., Minwoo, Y., JinSeok, P.: Intelligence in the Internet Era: understanding OSINT and case analysis (2013)
4. Wonhyung, P.: Security with open search service (OSINT) (2021)
5. Woong, C.: Open source intelligence in the information age (2008)
6. Javier, P.-G., Pantaleone, N., Félix, G. M., Gregorio, M. P.: The not yet exploited goldmine of OSINT: opportunities, open challenges and future trends (2016)
7. Irfan, S.: The art of searching for open source intelligence (2016)
8. Robert, A. F.: Open source intelligence methodology (2019)
9. Gajin, N., Neul, L.: A study on the limitations of OSINT and SOCMINT investigation and analysis work from a security perspective and the requirements for overcoming it (2021)
10. Yang, H. K., Lee, K. H., Choi, J. H.: A study on personal information hacking using domestic search engines (2007)

A Study on the Strategy of SWOT Extraction in the Metavers Platform Review Data: Using NLP Techniques

Jina Lee, Euntack Im, Inmo Yeo, and Gwangyong Gim

Abstract The Metaverse application industry has recently emerged as an important industry, and user review data has become a rich resource for business opportunities. This study proposes a framework that addresses the major shortcomings of traditional SWOT analysis through NLP based on review data from Metaverse applications. We collect review data on ZEPETO and ROBLOX from the Google Play Store and extract the aspects of each review data and perform sentiment analysis through natural language processing. Based on ZEPETO, we conduct research on the strategic positioning of Metaverse applications. The framework presented in this study presents ways to establish business strategic tools using secondary data and provides implications for important factors in building a Metaverse application environment.

Keywords Metaverse · SWOT analysis · Strategic positioning · Business straegy tool · Secondary data · Sentiment analysis · Application · NLP

1 Introduction

Recently Metaverse has emerged as an important issue in countries around the world. Metaverse has begun to attract attention as a new space that can maximize immersion as it coincides with a situation where non-face-to-face situations become

J. Lee · E. Im · G. Gim (✉)
Department of Business Administration, Soongsil University, Seoul, South Korea
e-mail: gygim@ssu.ac.kr

J. Lee
e-mail: kmtca1355@naver.com

E. Im
e-mail: iet030507@gmail.com

I. Yeo
Samsung Town Finance Center, Samsung Securities Co., Ltd., Seoul, South Korea
e-mail: visionfor1@naver.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_13

common and time and space problems need to be solved in all daily fields (education, games, work, consumption, etc.) because of Pandemic situation prolonged due to COVID- 19. Recently, global companies and governments have been aggressively investing and expanding in Metaverse, but empirical research and strategies how to use them are still insufficient. Various Metaverse-based businesses require research on strategic utilization plans and directions, and SWOT analysis has long been used as a proposed business strategy planning tool [1]. Successful strategies derived from SWOT analysis are based on suitability between internal competencies and external situations [2] and have been widely adopted throughout the industry due to the simplicity of processes that are easy to analyze and derive results [3]. However, SWOT analysis is too simple, leading to strategic planning errors [4] or mainly analyzed by subjective evaluation of experts, so it has a qualitative characteristic that additional quantitative analysis must be supplemented, and empirical verification is often insufficient [5].

Therefore, this study proposes a framework to solve the shortcomings of SWOT analysis while examining what is important in establishing a metaverse application environment by conducting SWOT analysis through machine learning techniques based on online review data of metaverse mobile applications (User Generated Contents). The SWOT analysis targets for this study are ZEPETO and ROBLOX, which currently provide metaverse mobile services, and through Google Play Store, reviews of US users using ZEPETO and ROBLOX are collected, doing sentimental analysis and aspect extraction, and data are placed on S, W, O, and T, respectively.

2 Theoretical Background

2.1 Metaverse

Metaverse is a combination of the words 'Meta' which means transcendence and 'Universe' and can be viewed as a virtual world or space connected to the physical world [6]. considers this to be a virtual mixture of digital and physical environments facilitated by convergence and expansion between the Internet and Web technologies and states that all individual users possess their respective avatars to utilize them similarly to the user's physical self or to experience alternative life. Various technologies such as blockchain (cryptocurrency, virtual assets, NFT, etc.), cloud, computer vision, artificial intelligence, and extended reality technology should be combined to build and utilize metaverse applications. From Google, Meta, and Microsoft, to companies in the fashion, entertainment, and financial industries, they have recently built or used metaverse applications. According to Grand View Research, the global metaverse market is expected to reach $678.8 billion by 2030 and to record a CAGR of 39.4% over the forecast period due to increased demand in the end-use industry [7]. Like the movie Ready Player One, Metaverse is still developing a wearable device that connects all parts of the body and allows you to experience a completely

new world. In addition, a metaverse application has emerged as a virtual world and a virtual space that can be immersed without immersive devices. These applications mainly provide users with various functions such as interaction between users such as SNS, playing games, and working in mobile and PC environments.

2.1.1 ZEPETO

Naver Z that aims to become a virtual world platform that anyone can create what they have dreamed of in their hearts based on 3D avatars and has more than 300 million users in more than 200 countries around the world. ZEPETO not only provides interaction between users and game functions by providing virtual spaces with various themes called 'world', but also provides a service that allows users to create virtual spaces on their own through its own program called 'Build it'. In addition, it provides a "ZEPETO Creator" function that allows users to produce various contents (short form contents, webtoons, etc.) based on avatars and through this function ZEPETO users can generates profits by producing and selling avatar items not only purchasing items through ZEPETO's virtual currency called "Zem" and "coin". ZEPETO provides a "LIVE" function that allows users to communicate with other users through real-time broadcasting using avatars and also collaborates with various brands. Collaboration is a virtual production of many products such as movies, cars, and fashion brands to provide users with various experiences, promote songs by various artists such as Black Pink and BTS, and hold virtual fan signing events to communicate with fans and artists' avatars. ZEPETO provides various functions based on avatars, especially communication and self-expression such as SNS, and is free to create and generate profits. In this study, ZEPETO is designated as its own company in SWOT analysis.

2.1.2 ROBLOX

ROBLOX is an online game platform and game creation system that explores millions of immersive 3D experiences created by the global developer community with the aim of allowing everyone to imagine, create, and have fun with their friends. ROBLOX is run by a global community of millions of developers who offer 'Roblox Studio' each month to produce their own immersive multiplayer experience and has 230 million users [8]. From the perspective of a metaverse, a virtual environment with a mix of physical and digital environments, ROBLOX can be viewed as a game-type metaverse application in terms of creating, selling, or purchasing games through its own virtual currency 'Robux', and providing a variety of games (virtual space or the world) for users to interact within ROBLOX [9]. ROBLOX, like ZEPETO, can use services in both mobile and PC environments, and is used as a virtual performance venue and educational platform such as collaboration with fashion brands such as VANS and NIKE, Virtual Grammy Week. In both applications, teenagers called Z generation, or zoomers, account for more than half of the users and can create and experience virtual spaces. In addition, it has something in common with ZEPETO that

it has its own virtual currency and communication and game functions. Therefore, in this study, ROBLOX was designated as a competitor for ZEPETO.

2.2 SWOT Analysis with Secondary Data

SWOT analysis identifies the strengths and weaknesses of the internal capacities of the target to be analyzed, and factors related to opportunities and threats in the external environment. This analysis is developed to eliminate weaknesses based on advantages and exploit opportunities or respond to threats. SWOT analysis has been widely adopted throughout the industry due to the simplicity of processes that facilitate analysis and decision-making [3] and is a strategic planning tool that can be used to achieve goals and efficiently serve customers [10].

There are still not many studies that have conducted SWOT analysis using secondary data. Most of these studies often semi-automated or automated SWOT analysis using machine learning techniques. In addition, the criteria for classifying S, W, O, and T are shown in Fig. 1. In the case of SWOT analysis using secondary data, the internal capability to classify strengths and weaknesses uses own company(or product)'s review data. The external environment, which classifies opportunities and threats, designates competitors and then classifies their strengths as their own threats and classify competitors' weaknesses as their own opportunities.

In the studies of [12, 13], both studies used review data to propose and verify the RE-SWOT MARTIX, that is, a semi-automatic SWOT analysis method. In the study of [11], feature extraction and sentiment analysis were performed using online review data in the tourism field, and each feature was assigned to S, W, O, and T based on the FPS score suggested in the study in [12]. FPS is a formula that categorizes S, W, O, T based on the number of reviews that mention product-related features and the user's rating-based feature words divided into positive and negative. In both studies, it is meaningful to semi-automate the SWOT analysis by classifying the user's ratings by performing sentiment analysis with stars(in reviews) or review texts, but there is a limitation that more research is needed on the verification of the FPS formula. In the study of [3], UGC (User Generated Contents) data including customer opinions and reviews was used to view and analyze the competitive advantage from the consumer's point of view. In this study, the disadvantages of the existing SWOT analysis are overcome by using the review data to automate the SWOT analysis using the NLP technique. In this study, aspect-based sentiment analysis is performed by collecting

Typical SWOT analysis

	Internal	External
Helpful	Strengths	Opportunities
Harmful	Weakness	Threats

SWOT Analysis with Secondary Data

	Own Product	Competing Product
Positive	Strengths	Opportunities
Negative	Weakness	Threats

Fig. 1 SWOT analysis classification

and pre-processing online reviews for tablet PCs of various brands on Amazon. In this case, the review data is collected by generations of products of several brands. After that, Aspect-sentiment pairs are created, Association Rules are extracted, and Aspect-sentiment pairs of the cluster of Association Rule mining are applied. Extract frequent association rule patterns based on different products of different product generations. Then, one brand product is designated as the own company, and if the difference from the previous generation within the same brand is positive, it is designated as a strength, and if it is negative, it is designated as a weakness. After extracting related rule patterns in the same way for reviews of products of other brands, researchers of this study look at those brands as competitors. Weaknesses of competitors (competitive brand products) are positioned as opportunities and strengths as threats. In the case of this study, since aspect-based sentiment analysis is performed and S and W are identified by product generation, there are limitations in that many analysis targets must exist in the market or labeled prior data is required.

3 Research Method

3.1 Research Model

In this study, metaverse mobile application ZEPETO is designated as 'own company (brand, product)' and ROBLOX is designated as 'competing company (brand, product)'. We conduct SWOT analysis with machine learning applied through the review data of both applications. First, we collected review data of ZEPETO and ROBLOX from the Google Play Store. The collection period is from January 1, 2021 to March 17, 2022, and the data collected thereafter is separated into sentences. The separated sentence data was subjected to sentiment analysis through the VADER algorithm to derive a polarity value (compound). In addition, the separated sentence data was pre-processed through tokenization, part-of-speech tagging, and stop word removal, and then aspects of each application were extracted and classified through LDA topic modeling. Afterwards, the aspect corresponding to each sentence and the derived polarity value (compound) are composed into one dataset, and the average of the polarity values (compound) of each aspect is calculated as a positive aspect if it exceeds 0, and a negative aspect if it is less than 0 classified. Finally, positive aspects of ZEPETO were placed in S, negative aspects were placed in W, positive aspects of ROBLOX were placed in T, and negative aspects were placed in O, and the results were interpreted. The research model of this study is shown in Fig. 2.

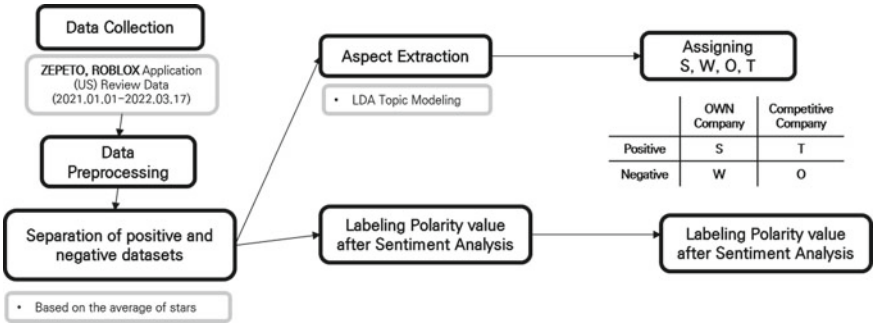


Fig. 2 Research model of this study

3.2 Research Method

3.2.1 Data Collection and Preprocessing

The data to be used in this study is the review data left by users after using the application directly, and was collected after writing the code using Python in the Google Colab Pro environment. The data collected reviews from US users on the Google Play Store, and the collection period is from January 1, 2021 to March 17, 2022. The number of collected review data was 126,824 in ZEPETO and 82,784 in ROBLOX, respectively, and each collected data was first separated into sentences. As for the data of the separated sentence unit, ZEPETO had 306,666 of ZEPETO and ROBLOX had 332,436. Thereafter, positive datasets and negative datasets were constructed based on the average of the star (Out of five) of each application review data. To extract aspects, only features were extracted and tokenization was performed, and part of speech was tagged to each feature to extract only nouns and verbs and remove stop words. At this time, exclamations classified as abbreviations or nouns were produced as a stop word dictionary and removed, and characters of less than 2 characters were deleted.

3.2.2 Extracting and Classifying Aspects

In this study, LDA Topic Modeling was performed on preprocessed review data to extract aspects. There are no classification standard for S, W, O, T of the metaverse application or a set evaluation standard. In this background, there was no labeled data, so LDA Topic Modeling, an unsupervised learning method, was used. As mentioned above, the latest industrial applications that do not have criteria to classify topics are the subject of this analysis and as mentioned in Chap. 2, although ROBLOX and ZEPETO have a lot in common, each has different detailed functions and their own terminology. So, the extracted topics (aspects) were arbitrarily designated by the researcher based on the classified keywords. Therefore, in this study, research

and previous studies on ROBLOX and ZEPETO were reviewed, and topics were designated based on the experience of using each application for about two months.

3.2.3 Sensitivity Analysis and Dataset Composition

Sentiment analysis was performed on the sentence data of each review data separated based on the average of the stars in the review, and in this case, the rule-based VADER algorithm was used. The reason for using VADER is that this algorithm performs well when dealing with texts from social media, movie reviews, and product reviews [13]. The VADER algorithm gives a sensitivity score between −1 and 1, and a 'compound ' is derived by appropriately combining positive, negative, and neutral values. In this study, the compound was used as the polarity value.

3.2.4 Assign Dataset to SWOT

A final dataset was constructed based on the classification results of aspects extracted through sentiment analysis and LDA topic modeling. Based on the sentences separated from each positive and negative review dataset, the dataset consists of the emotional polarity value for each sentence and the topics with the highest weight in the sentence. Afterwards, in this study, since ZEPETO was designated as own company, positive aspects of ZEPETO were placed in S and negative aspects of ZEPETO were placed in W. Positive aspects of ROBLOX are placed in T and negative aspects are placed in O. The average of the polarity values of each sentence was calculated based on the topic, and the importance was judged by the ranking of the topics arranged in each S, W, O, and T based on this.

4 Results

In this study, SWOT analysis was conducted by designating ZEPETO as the own company and ROBLOX as the competitor. As for the data set used at this time, according to the criteria of previous studies, the positive dataset of ZEPETO was used for S (strengths) and the negative dataset of ZEPETO was used for W (weakness). The negative dataset of ROBLOX was used for O(opportunity) and the positive dataset of ROBLOX was used for T(threatens). Afterwards, LDA topic modeling was carried out based on each data set, and the topics(aspects) were arranged on S, W, O, T. The results were interpreted by judging the importance of each aspect by taking the average of the sentiment analysis results (polar value; compound) of aspects.

In addition, in this study, negative aspects of ROBLOX can be viewed as opportunities for ZEPETO, but overlapping topics (aspects) were judged as common challenges of metaverse applications. However, the aspects of each application were identified and interpreted through the keywords of each aspect. Similarly, positive aspects of

Table 1 SWOT analysis result

	ZEPETO (own)	ROBLOX (competitor)
Positive	**S** 1. Creating items 2. Character(Avatar) 3. Feeds (Avatar-based Shortform Content) 4. K Pop 5. Playing Game 6. Communication (Interaction) 7. Worlds(virtual spaces)	**T** 1. Connecting devices 2. Experience 3. Making game 4. Communication (Interaction) 5. Playing game 6. Character(avatar)
Negative	**W** 1. Inducing excessive spending (zem) 2. Security 3. Meeting Strangers 4. Hacking 5. System Error	**O** 1. Inappropriate contents 2. Inappropriate meeting 3. System error 4. Hacking and fraud 5. Addiction

ROBLOX can be seen as threats of ZEPETO, but overlapping topics (aspects) were judged as common advantages of metaverse elements, and aspects of each application were identified and interpreted through the order and keywords of each aspect. The final SWOT analysis results are shown in Table 1.

4.1 SWOT: S, W (ZEPETO Positive and Negative Data)

Each aspect was placed in S (strengths) through the positive dataset of ZEPETO separated based on stars. Table 2 shows the average and rank of each derived aspect, keyword, and polarity value. Aspects are Communication (Interaction), Game, Character (Avatar), Feeds (Avatar-based Shortform Content), Creating items, Worlds (virtual spaces), K Pop. Aspects are specified based on keywords derived through LDA topic modeling, and these constitute S. In addition, as a result of analyzing the importance of each aspect in S based on the average of the polarity values, Creating items>Character (Avatar)>Feeds (Avatar-based Shortform Content)>K Pop>Game>Communication (Interaction)>Worlds (virtual spaces) is the order. In this case, the closer to 1, the higher the priority because the positive dataset was used.

First, as a result of averaging the polarity values of aspects placed in S, aspects related to avatars were selected in the 1st and 2nd rankings for ZEPETO. ZEPETO allows users to use it as a multi-persona through 3D-based avatars and various items, and can create avatars similar to users with user photos through its own technology. In addition, it is interpreted that this result was obtained because it is providing a variety of items through collaboration with various brands. In addition, ZEPETO

Table 2 S(Strengths) and W(Weakness) aspects results

S: ZEPETO Positive				W: ZEPETO Negaitive			
Importance ranking	Topic	Keyword		Importance ranking	Topic	Keyword	
6	Communication (Interaction)	Make	Friend	5	System error	Bugs	Crashing
		Friends	Hangout			Laggy	Loading
		Meet	Follow			Uninstall	Post
		Chat	Talk			Disconnection	Glitches
		People	Play			System	Download
5	Playing game	Game	Minigame	1	Inducing excessive spending (zem)	Zem	Purchase
		Player	Download			Induce	Ads
		Group	Update			Coin	Diamond
		Gameplay	Laggy			Expensive	Quest
		Jumpmaster	Party			Items	Clothes
2	Character (Avatar)	Character	Hair	4	Hacking	Hacking	Error
		Outfits	Clothes			Account	Login
		Fashion	Item			Hacker	Problem
		Buy	Change			Email	Delete
		Edit	Price			Password	Games
3	Feeds (Avatar-based Shortform Content)	Videos	Dances	2	Security	Stalking	Spying
		Make	Photos			Track	Spyware
		Entertainment	Watch			Avatar	Virus
		Tiktok	Followers			Camera	Records
		Followers	Challenges			Photos	Trackers

(continued)

Table 2 (continued)

S: ZEPETO Positive				W: ZEPETO Negaitive			
Importance ranking	Topic	Keyword		Importance ranking	Topic	Keyword	
1	Creating items	Avatar	Item	3	Meeting Strangers	Scam	Meeting
		Create	Make			Kids	Kidnaps
		Edit	Creator			Stranger	Chat
		Clothes	Zepetostudio			Dating	Crews
		Zem	Earn			Child	Scammers
7	Worlds (Virtual spaces)	World	Experience	**BTS member name			
		Make	Map				
		Graphics	Theme				
		Epic	Loading				
		Graphics	Creator				
4	K Pop	BTS	Twice				
		Blackpink	Jungkook				
		Entertainment	Jimin**				
		Fans	Idol				
		K Pop	Pop				

allows users to create various items by using 3d design tools directly from simple image upload through the mobile application (Becoming a Creator) and PC (ZEPETO studio). Users can also sell the items to generate revenue through ZEPETO's virtual currency called zem. The fact that you can make a variety of items yourself very easily and sell them to generate profits is interpreted as being recognized as a very big advantage for ZEPETO users. 'Feeds', which can produce and view short-form content and images (photos) based on the user's avatar, also ranked third in ZEPETO's strengths.

Many users of ZEPETO are Generation Z, and they enjoy short video content such as tiktok and shorts. From this point of view, it is interpreted that it was selected as a strength because it is possible to easily create various contents based on the avatar. In addition, K-pop has recently been gaining popularity in global countries, and K-pop artists, who had difficulty communicating with their fans due to the coronavirus, conducted various virtual spaces, avatars, and events through collaboration with ZEPETO. Accordingly, fans of Kpop artists encountered various Kpop contents through ZEPETO, and from this point of view, it is interpreted that Kpop is placed as ZEPETO's strength. In addition, the world and game that function as virtual space, which can be seen as the biggest characteristics of the metaverse, were selected as strengths.

Table 2 also shows the averages and ranks of aspects, keywords, and polarity values arranged in W (weakness). Aspects are System Error, inducement to pay (zem), Hacking, Security, and Meeting Strangers and as a result of analyzing the importance of W of each aspect based on the average of the polarity values, the order is Inducing excessive spending (zem)>Security>Meeting Strangers>Hacking>System Error. In this case, the closer to -1, the higher the priority because a negative data set is used. As a result of the average polarity value of aspects placed in W, the first is about paying zem, the ZEPETO virtual currency. In ZEPETO, zem and coin are used to purchase various items, and items are sold through collaboration with many brands. Users can use the avatar as a multi-persona and purchase products from expensive brands as avatar items at a much lower price in reality. From this point of view, it is interpreted that users spend excessively on item purchases and are placed on a weakness. Also, in the case of security, which is second, if you look at keywords, you can see words such as tracking, stalking, and records. As a result of the investigation in this regard, there was a content that the photos used to generate avatars in ZEPETO and the user's activity records (log records) could be used or recorded regardless of the user's intention. However, this part was collected with the consent of the user, and most of the reviews related to this were confirmed to be rumors. Third was about ethical problems that could arise from meeting various people in an avatar-based virtual space where anonymity was guaranteed. In particular, users can meet various people through ZEPETO's 'crew' function and 'match' function, but from a different point of view, you can meet strangers and be exposed to crimes such as online grooming. 4th and 5th places were also related to system errors such as delays, errors, disconnections, and crashing, as well as problems caused by account hacking.

4.1.1 SWOT: O, T (ROBLOX Positive and Negative Data)

Table 3 is the result of placing each aspect in O (Opportunity) through the negative dataset of ROBLOX. Aspects of O are Addiction, System error, Hacking and fraud, Inappropriate contents, Inappropriate contents, and Inappropriate meeting and the order of importance is Inappropriate contents>Inappropriate meeting>Making game>System error>Hacking and fraud>Addiction. First, in the case of 'inappropriate contents', it is about ethical issues that can occur in an anonymous virtual space. There are various game types such as survival, adventure, and action in ROBLOX, and there are many games that allow actions such as killing or attacking a certain target. In the virtual space where immersion is maximized, these games can experience inappropriate content and furthermore, by playing certain roles and actions, can instill awareness of unethical behavior. This is a part that has been frequently mentioned as one of the negative problems of online games. In addition, there were a lot of reviews saying that there are virtual spaces where you can experience unethical and sexual content like 'condos' in ROBLOX, allowing young users to provide inappropriate content in many ROBLOX. Here, 'condos' is a term commonly used when referring to Roblox sex games, and it is a space created by users, where people can talk about sex and their avatars can have virtual sex [14].

This is interpreted as something that ZEPETO can take as an opportunity by being careful and preventing it in advance when developing the game. "addiction" means that users are addicted to games, accessing and playing for a long time, or recklessly paying for ROBLOX's cryptocurrency, "robux". The addiction problem is one of the issues of ROBLOX, and it is interpreted as highly addictive because most users are young or can play many kinds of games with various devices. In addition, in the case of ROBLOX, it is interpreted that many users can spend indiscriminately because a significant part of it is a paid game, that is, because they have to pay to play the game. In the case of ZEPETO, it is interpreted that this can be an opportunity for ZEPETO, as it does not require extra spending to enjoy the game except for decorating the current avatar. In this regard, ROBLOX provides monthly spending restrictions, spending notification control functions, and game account restrictions to help parents prevent their children from excessive spending, game addiction, and inappropriate experiences.

Table 3 shows the results placed in T (threatens) through the positive data set of ROBLOX. Aspects are Experience, Playing game, Character(avatar), Connecting devices, Interaction, Making game. As a result of analyzing the T priority of each aspect, the order is Connecting devices>Experience>Making game>Interaction>Playing game>Character (avatar). First, connecting devices is ZEPETO's number one threat, and both applications provide pc and mobile versions, but in the case of ROBLOX, it provides services through a game-only device called xbox to induce more immersion and interest. In addition, ROBLOX supports Oculus Rift and HTC Vive devices to provide a more immersive experience, and global major companies are entering the VR and AR device market recently. Against this background, in the case of ZEPETO, it is interpreted that it is necessary to provide the experience that ROBLOX provides through support for more diverse devices. In

Table 3 W(Weakness) and T(Treathens) aspects results

T: ROBLOX Positive				O: ROBLOX Negaitive			
Importance ranking	Topic (Aspect)	Keyword		Importance ranking	Topic (Aspect)	Keyword	
2	Experience	Role	Bloxburg	5	Addiction	Games	Dollars
		Simulator	Livetopia			Playing	Robux
		Family	Meepcity			Addicted	Purchase
		Cousin	Brookhaven			Hours	Stuck
		Adopt	Tycoons			Spending	Items
5	Playing game	Games	Connect	3	System error	Fix	Disconnection
		Playing	Adventures			Server	Update
		Obbys	Simulator			Glithces	Fail
		Roleplay	Multiplayer			Lag	Reconnect
		Noob	Mode			Bugs	Login
6	Character(avatar)	Character	Avatars	4	Hacking and fraud	Hacking	Password
		Clothes	Buy			Scammers	Account
		Bacon	Customize			Exploiter	Robux
		Hair	Style			Cheater	Trading
		Acc	Premium			Scam	Player
1	Connecting devices	Tablet	Mobile	1	Inappropriate contents	Slender	Problems
		Xbox	Connection			Killing	Murder
		Phone	Laptop			Condos	Bypass
		Lagging	Device			Adults	Toxic
		iPad	Problem			Sex	Violence

(continued)

Table 3 (continued)

T: ROBLOX Positive				O: ROBLOX Negaitive			
Importance ranking	Topic (Aspect)	Keyword		Importance ranking	Topic (Aspect)	Keyword	
4	Communication (Interaction)	Friends	Games	2	Inappropriate meeting	Chat	Community
		Make	Interact			People	Couple
		Community	Teens			Slender	Sexy
		Play	People			Bullying	Moderator
		Meet	Talk			Dating	Messages
3	Making game	Game	Robux				
		Tycoons	Money				
		Make	Creator				
		Code	Developer				

addition, ROBLOX provides games that can be experienced in a variety of ways, such as role-playing and simulator. These games enable users to create their own business like playing 'tycoon', from simply experiencing a certain role. 'tycoon' is a genre of experience in ROBLOX where the player owns his or her base (usually a business or company) [14]. As such, it is interpreted that ZEPETO needs to provide a function that enables a more diverse experience within the virtual space. ROBLOX also has the ability to easily create games and provides a ROBLOX studio where users can design and create their own games, environments and other objects through a simple coding language called LUA [15]. In this regard, ROBLOX operate the 'Roblox Star Program', an invitation-only program for top developers who want to expand and expand the license of intellectual property (IP) related to the ROBLOX experience to ROBLOX. The Roblox Star Program is a program designed to grant Roblox an exclusive worldwide license to use the developer's username, avatar, and any user-generated content they choose to include [16]. This enables ROBLOX to create games and content, monetize it and attract more users [17]. ZEPETO also provides tools to create virtual spaces, but currently, virtual spaces containing game functions can be created through 'unity' by writing a Typescript. However, since the feature was released not long ago, many users are not actively using it. From this point of view, ZEPETO is expected to be better if it provides tools and functions that can directly create various games through simple and easy coding and programming like ROBLOX.

5 Conclusions

In this study, UGC(User Generated Contents) was used, but since data from a new industry was used, there was no labeled data or no evaluation criteria, so subjectivity could not be completely excluded. For this, feature extraction using more diverse unsupervised learning methods is required. Therefore, it is necessary to conduct comparative analysis through reviews of various metaverse applications other than ZEPETO and ROBLOX, or to supplement the limitations by using more diverse data such as article data in addition to online reviews. In addition, even though aspects were placed in S, W, O, and T, simple noun phrases and noun-centered aspects were used, so there is a limitation that human subjective judgment and interpretation are required for strategic use. However, this study was able to overcome the limitations that the existing SWOT analysis mainly consisted of subjective evaluation by experts or there was no empirical verification by presenting a framework for deriving business strategies through online review data. It is also expected that this framework will be useful in industries with evaluation criteria or labeled data. It is meaningful to provide practical implications for building various metaverse applications and providing services in the coming metaverse era by understanding what characteristics are important for building metaverse applications, which are emerging as an important industry recently.

References

1. Niranjanamurthy, M., Nithya, B.N., Jagannatha, S.J.C.C.: Analysis of blockchain technology: pros, cons and SWOT. Clust. Comput. **22**(6), 14743–14757 (2019)
2. Agarwal, R., Grassl, W, Pahl, J.: Meta-SWOT: Introducing a new strategic planning tool. J. Bus. Strat. (2012)
3. Cheng, L.C., et al.: User-defined SWOT analysis–A change mining perspective on user-generated content. Inf. Process Manag. **58.5**,102613 (2021)
4. Pickton, D.W., Weight, S.: What's swot in strategic analysis? Strat. Chang. **7**(2), 101−109 (1998)
5. Gurl, E.: SWOT Analysis: A Theoretical Review (2017)
6. Lee, L.H et al.: All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research agenda (2021). arXiv preprint arXiv:2110.05352
7. Bloomberg: Metaverse Market Size Worth $678.8 Billion by 2030, Grand View Research, Inc. https://www.bloomberg.com/press-releases/2022-03-09/metaverse-market-size-worth-678-8-billion-by-2030-grand-view-research-inc. (2022)
8. ActivePlayer.io: Roblox Live Player Count and Statistics. https://activeplayer.io/roblox/#:~:text=Roblox%20has%20over%20230%20Million%20registered%20players%20and%2030%20million%20daily%20players
9. Santoro, J: Does Roblox Have a Plan to Win the Metaverse?. https://www.fool.com/investing/2022/02/12/does-roblox-have-a-plan-to-win-the-metaverse/ (2022)
10. Culp III, K., et al.: Using a SWOT Analysis: Taking a Look at Your Organization (2016)
11. Fehringer, D.: Six steps to better SWOTs. Compet. Intell. Mag. **10**(1), 54 (2007)
12. Dalpiaz, F., Parente, M.: RE-SWOT: from user feedback to requirements via competitor analysis. In: International Working Conference On Requirements Engineering: Foundation For Software Quality, pp. 55−70. Springer, Cham (2019)
13. Tu, S.F., Hsu, C.S., Lu, Y.T.: Improving RE-SWOT analysis with sentiment classification: a case study of travel agencies. Futur. Internet. **13** (9), 26 (2021)
14. Clayton, A., Dyer, J.: Roblox: The children's game with a sex problem. BBC News, 2022.02.15. https://www.bbc.com/news/technology-60314572#:~:text=Roblox%20sex%20games%20are%20commonly,thrown%20out%20of%20the%20window
15. Tycoon: Roblox Wiki. https://roblox.fandom.com/wiki/Tycoon
16. Roblox Star Program: Devleoper, https://developer.roblox.com/en-us/resources/Stars-About#:~:text=The%20Stars%20program%20is%20designed,the%20ones%20they%20don't
17. Virginia: Can roblox be educational? absolutely! Here's how.., iD Tech, 2021.05.24, https://www.idtech.com/blog/roblox-educational-benefits

Does Facial Expression Accurately Reveal True Emotion? Evidence from EEG Signal

Huy Tung Phuong, Yangyoung Kun, Jisook Kim, and Gwangyong Gim

Abstract Facial expression is one of the ways in which human expresses emotion. However, there has been evidence that facial expression does not always accurately reflect inner emotions, leading to the shaking of theories about human basic universal emotions. In this study, we conduct an experiment to collect facial expression data and EEG signals of participant. The results indicate that not all kinds of basic emotion were expressed in facial expressions under the experiment environment. On the other hand, brain signals have proven to be a highly reliable tool for emotion recognition task. This study shows the problems faced by facial emotion recognition systems and proposes future works to improve the efficiency of those systems.

Keywords EEG · Facial expression · Emotion measurement · Emotional AI · Affective computing

1 Introduction

Affective computing has emerged as an important field of study that aims to develop systems that can automatically recognize emotions. Emotion recognition has gained substantial attention in the last decade because it is directly linked to psychology,

H. T. Phuong · G. Gim (✉)
Department of Business Administration, Soongsil University, Seoul, South Korea
e-mail: gygim@ssu.ac.kr

H. T. Phuong
e-mail: tungph@soongsil.ac.kr

Y. Kun
Department of IT Policy and Management, Soongsil University, Seoul, South Korea
e-mail: yyk0918@naver.com

J. Kim
Management Planning HQ, Soongsil University, Seoul, South Korea
e-mail: inispriblue@gmail.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_14

physiology, learning studies, marketing, and healing. In 2019, the market for emotion-detection technology is worth roughly 21.6 billion dollars, and its value is predicted to reaching 56 billion dollars by 2024 [1].

Emotion has been defined in various ways in the twentieth century, and there is currently no scientific consensus on a definition [2, 3]. From neurophysiological perspective, emotions are mental states brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure [3–7]. From mechanistic perspective, in the meanwhile, emotions can be defined as a positive or negative experience that is associated with a particular pattern of physiological activity [8, 9].

Facial expression is a convenient way for humans to communicate emotion. As a result, research on emotional facial expression recognition has become a key focus area of personalized human−computer interaction [10–12]. Influential observations in the 1960s and 1970s by US psychologist Paul Ekman suggested that, around the world, humans could reliably infer emotional states from expressions on faces−implying that emotional expressions are universal [13, 14]. This assumption influences legal judgments, policy decisions, national security protocols, and educational practices; guides the diagnosis and treatment of psychiatric illness, as well as the development of commercial applications; and pervades everyday social interactions as well as research in other scientific fields such as artificial intelligence, neuroscience, and computer vision [15].

These ideas stood largely unchallenged for a generation. But a new cohort of psychologists and cognitive scientists has been revisiting those data and questioning the conclusions. Many researchers now think that the picture is a lot more complicated, and that facial expressions vary widely between contexts and cultures [16].

Although AI companies market software for recognizing emotions in faces, psychologists still debate whether expressions can be read so easily [16]. Facial expressions are extremely difficult to interpret, even for people. With researchers still wrangling over whether people can produce or perceive emotional expressions with fidelity, many in the field think efforts to get computers to do it automatically are premature−especially when the technology could have damaging repercussions [16]. Thus, during the last decade, many research and development efforts have been deployed to develop new approaches and techniques for emotion recognition. It is becoming increasingly attractive to detect human emotions using biological brain signals. Electroencephalography (EEG) is a reliable and cost-effective technology used to measure brain activity [1].

In this study, we re-examine the results of recent studies, that facial expressions do not completely accurately represent real emotions [16]. We also test our hypothesis that EEG signals would provide a better tool to measure emotions more accurately than facial expressions. In our experiment, participant was evoked to Ekman's six basic emotions (sad, surprise, enjoy, disgust, anger, fear) [17] through videos played on a computer screen. While the participant watched the emotional videos, EEG signals were recorded in 14 channels of signal. At the same time, the participant' facial expressions were recorded by a camera mounted above the computer screen.

2 Theoretical Background

2.1 Emotions as Discrete Categories: Basic Emotion Theory, Emotional Facial Expression, and Relevant Debates

Basic emotion theory proposes that human beings have a limited number of emotions (e.g., fear, anger, joy, sadness) that are biologically and psychologically "basic", each manifested in an organized recurring pattern of associated behavioral components [18].

This idea of basic emotion is not new. For instance, in the seventeenth century, Descartes (1649/1996) already identified six "primitive passions" ("passion" formerly being used as a term for "emotion"): wonder (admiration), desire, love, joy, hatred, and sadness [16]. Paul Ekman [13, 14, 17] identified the six basic emotions as anger, surprise, disgust, enjoyment, fear, and sadness. Ekman suggested that, around the world, humans could reliably infer emotional states from expressions on faces—implying that emotional expressions are universal. His studies indicated that some expressions, and their corresponding feelings, were recognized by people of all cultures.

The viewpoint of basic emotion theory has had detractors. Many researchers now think that the picture is a lot more complicated, and that facial expressions vary widely between contexts and cultures, thus researchers are increasingly split over the validity of Ekman's conclusions [16].

The observational study of Chen et al. [19] reported that people experiencing pain or orgasm produce facial expressions that are indistinguishable, which questions their role as an effective tool for communication. This study found that although Westerners and East Asians had similar concepts of how faces display pain, they had different ideas about expressions of pleasure.

According to Barret et al. [15], the available scientific evidence suggests that people do sometimes smile when happy, frown when sad, scowl when angry, and so on, as proposed by the common view, more than what would be expected by chance. However, how people communicate anger, disgust, fear, happiness, sadness, and surprise varies substantially across cultures, situations, and even across people within a single situation. The study concluded that, there was little to no evidence that people can reliably infer someone else's emotional state from a set of facial movements.

A study conducted by Crivelli et al. [20], with two experiments on how residents of Papua New Guinea interpret facial expressions produced spontaneously by other residents of Papua New Guinea, also showed similar results. In both experiments of Crivelli et al., agreement with Ekman's predicted labels was extremely low. The study found no evidence for Ekman's conclusions, thus concluded that "trying to assess internal mental states from external markers is like trying to measure mass in metres" [16, 20].

Our research supports the view of Crivelli [20], Chen [19], Barret [15] and Heaven [16] on the assertion that facial expressions are complex and cannot be an accurate measurement of emotion. Thus, our view pointed to the reliability problem for emotional AI researchers who are training their artificial intelligence using classic dataset of faces for recognizing emotions.

2.2 Emotion as Dimensional Model: The Valence—Arousal Circumplex Model

Clinicians and researchers have long noted the difficulty that people have in assessing, discerning, and describing their own emotions [21]. This difficulty suggests that individuals do not experience, or recognize, emotions as isolated, discrete entities, but that they rather recognize emotions as ambiguous and overlapping experiences. Emotions seem to lack the discrete borders that would clearly differentiate one emotion from another, thus discrete emotional classes (i.e., Ekman's six basic emotions) are not representative of the full spectrum of emotions displayed by humans on a daily basis [22].

Indeed, researchers exploring the subjective experience of emotion have noted that emotions are highly intercorrelated both within and between the subjects reporting them [23, 24]. Subjects rarely describe feeling a specific positive emotion without also claiming to feel other positive emotions [24]. These intercorrelations among emotions, often obscured in experimental paradigms of basic emotions, are addressed head-on by dimensional models of affect. Dimensional models regard affective experiences as a continuum of highly interrelated and often ambiguous states [21].

The circumplex model of emotion was developed by James Russell in 1980 [25]. This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical axis and valence represents the horizontal axis, while the center of the circle represents a neutral valence and a medium level of arousal [26]. In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both factors. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and affective states [27].

An attractive feature of dimensional approaches is their parsimony and their applicability across multiple domains. Dimensional approaches have also proven to be empirically powerful, successfully accounting for a wide range of emotion effects [28].

2.3 Emotion Measurement Using Frontal EEG Asymmetry

Electroencephalography (EEG) is one of popular physiological indicators of emotion, beside skin conductance, pupil dilation, heart rate, and functional magnetic resonance imaging (fMRI) [2]. It is a reliable and cost-effective technology used to measure brain activity [1].

Brain waves are oscillating electrical voltages in the brain measuring just a few millionths of a volt. According to [29], there are five widely recognized brain waves, and the main frequencies of human EEG waves are: gamma (>35 Hz); beta (12−35 Hz); alpha (8−12 Hz); theta (4−8 Hz), delta (0.5−4 Hz).

Detecting emotion using EEG signals involves multiple steps being performed in sequence to satisfy the requirements of a brain–computer interface (BCI). Traditionally, these steps include removing artifacts from EEG signals, extracting temporal or spectral features from the EEG signal's time or frequency domain, respectively, and finally, designing a multi-class classification strategy. Feature quality dramatically increases the accuracy of the emotion classification strategy [1]. The most popular features in the context of emotion recognition from EEG are band power features from different frequency bands (Table 1). This assumes stationarity of the signal for the duration of a trial. Commonly used algorithm to extract the band power features are the estimation of Power Spectral Density (PSD) [30]. Welch [31] described the method to calculate PSD using fast Fourier transform in his publication, called Welch's method.

Literature studies pointed specifically to the alpha and beta frequency bands for emotion recognitions [32]. Schubring and Schupp [33], by conducting two experiments on 16−18 participant, have confirmed that high arousal is associated with talking in alpha and lower beta band power. Beta waves are associated with a state of an alert or excited mind, while alpha waves are more dominant in a state of relaxation.

Location in the brain is associated with different functions of brain [34]. Figure 1 illustrates the four main lobes of the human brain namely frontal lobe (including prefrontal lobe part), parietal lobe, occipital lobe, and temporal lobe. A parietal lobe of brain takes care of processing of nerve impulses related to senses and language functions, and a temporal lobe is a primary organization of sensory input [35]. The prefrontal lobe of brain, according to Blaiech et al. [36], plays a crucial role in the regulation of emotions and conscious experience. Inactivation of the left frontal

Table 1 Five basic EEG bands

Frequency band	Frequency	Brain states
Gamma	>35 Hz	Concentration
Beta	12−35 Hz	Anxiety dominant, active, external attention, relaxed
Alpha	8−12 Hz	Very relaxed, passive attention
Theta	4−8 Hz	Deeply relaxed, inward focused
Delta	0.5−4 Hz	Sleep

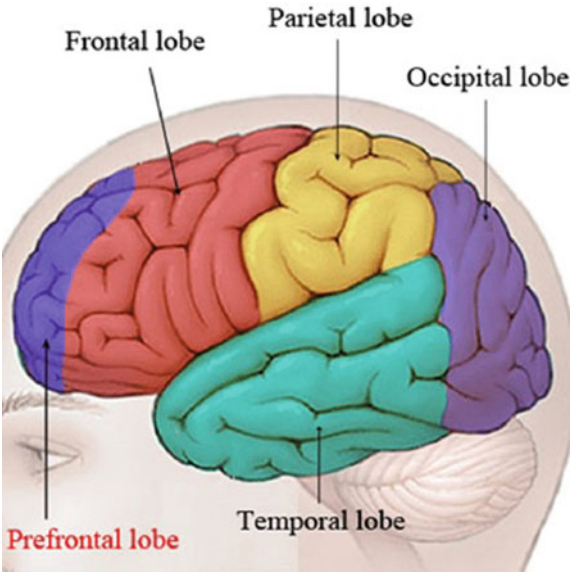


Fig. 1 Human brain structure [38]

indicates a negative emotion, and even inactivation of the right frontal indicates a positive emotion. Thus, the frontal EEG asymmetry's beta/alpha ratio is a reasonable indicator of the emotional state of a person [36, 37].

3 Methods

3.1 Participant and Experimental Procedure

In our study, data from one participant (female; age 24) was recorded. The participant had no history of any neurological, psychiatric, or hearing problems. The participant gave written informed consent before the experiment. The experiment was performed in accordance with the Declaration of Helsinki [40].

We used a subset of 30 video clips to evoke participant' emotion. We selected these videos from a database of 2,185 emotional video clips by Cowen and Keltner [41]. We focused on six categories of videos to evoke Ekman's six basic emotions (e.g., sad, surprise, enjoy, disgust, anger, fear). During the experiment, the participant' facial expressions were recorded by a camera placed above the computer screen.

At the beginning of the experiment, we explained the purpose of the study and the procedure for conducting the experiment. Then we set up the EEG device on the participant' head. After we had a stable EEG signal, we showed the participant 6 groups of videos that we had selected. After each group of videos, there was a rest time for the participant to relax and report the emotions they felt while watching the videos.

3.2 Data Recording and Preprocessing

The Emotiv EPOC X neuroheadset (a product of Emotiv Inc) [42] was used to obtain the necessary EEG raw data signals. The neuroheadset EPOC X is an EEG device that benefits from the ability of the nervous tissue to generate quantifiable electric potentials to measure brain activity. The electrodes are metallic with a plastic base and are placed on the scalp to measure brain activity. For optimal results, the electrodes must be wet with a few drops of a saline conductive solution to increase conductivity, and to improve their performance [39].

The headset is equipped with a total of 14 electrodes to measure brain activity, plus two extra electrodes that work as reference points. The electrodes are distributed on key points of the scalp as suggested by the international system 10–20 [39] as shown in Fig. 2. The 14 electrode positions used by the EPOC are labeled as AF3, F7, F3, FC5, T7, P7, 01, 02, P8, T8, FC6, F4, F8, and AF4.

In this study, the EEG signal of all 14 channels was recorded in 256 Hz sampling mode, and the raw time series data was stored by EmotivPro software as a CSV file. Simultaneously, facial expression was recorded by a high-definition camera and saved as video in MP4 format.

After obtaining the raw data, we used EEGLAB software [43] to preprocess the data, including artifact removing. Artifact is any component of the EEG signal that is not directly produced by human brain activity. EEG artifacts can be classified depending on their origin, which can be physiological or external to the human body (non-physiological). In our study, we used MARA (Multiple Artifact Rejection Algorithm) plugin for artifact rejection. The core of MARA is a supervised machine learning algorithm that learns from expert ratings of 1290 components by extracting

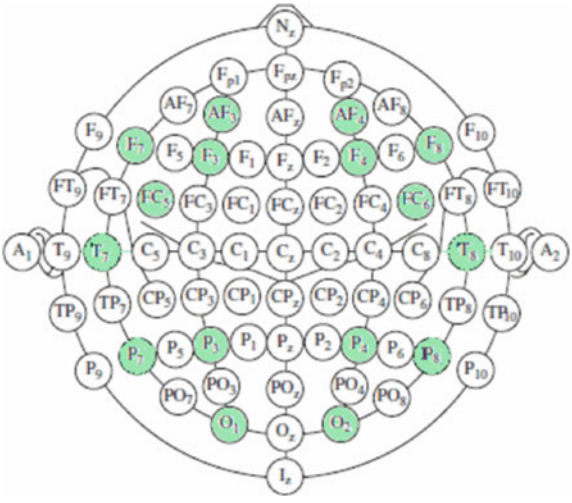


Fig. 2 EEG channels' location in the brain suggested by the international system 10–20 [39]

Table 2 Caculating valence and arousal

Valence	**Argument**: Positive, happy emotions result in a higher frontal coherence in alpha, and higher right parietal beta power, compared to negative emotion [34, 36, 37]. **Equation**: $Valence = \alpha F4/\beta F4 - \alpha F3/\beta F3$ (*)
Arousal	**Argument**: Excitation presented a higher beta power and coherence in the parietal lobe, plus lower alpha activity [34, 36, 37]. **Equation**: $Arousal = \frac{\alpha(AF3+AF4+F3+F4)}{\beta(AF3+AF4+F3+F4)}$ (*)

(*) Where αi and βi = PSD of alpha and beta frequency range obtained from *ith* channel of the EEG signal

six features from the spatial, the spectral and the temporal domain. Features were optimized to solve the binary classification problem "reject vs. accept". Thus, MARA should be able to handle eye artifacts, muscular artifacts, and loose electrodes equally well [44].

3.3 Data Analysis

The emotions that participant reported after watching each group of videos was used to evaluate the validity of those videos. We assumed that if the reported emotion matches the emotion the video attempts to elicit, the data collected is valid.

As mentioned in Chap. 2, brain activation by changes in emotion are mostly captured by electrodes located in the frontal area, e.g., AF3, AF4, F3, and F4 [34]. In this study, we extract the mean PSD feature from alpha band and beta band of these channels using Welch's method. Then, we calculated the value of valence and arousal based on the equations mentioned in Table 2. We then represented these emotional states in the valence−arousal two-dimensional space (the circumplex model).

For facial expression data, we performed a blind analysis of facial expressions with the help of an expert in this field. The expert received the 12 recorded videos of participant expression (6 videos for each participant) without sound and did not know the sequence in which emotions were evoked. The expert was asked to rate and point out which is the main emotion of participant' facial expression in each video.

4 Results

The emotions that participant reported after watching each group of videos was used to evaluate the validity of those videos. The results showed that all the reports that the participant gave were consistent with the videos' emotions. Thus, the obtained data is valid.

Table 3 The mean PSD (sample for sadness emotion). (Due to the limited number of pages, we only present the case of sad emotion)

Emotion	Mean PSD	Channel			
		F3	F4	AF3	AF4
Sad	meanPSD_alpha	2.6205	5.4297	0.3173	3.3108
	meanPSD_beta	−1.2412	1.7195	−3.4945	−0.3623

Table 4 Valence—arousal calculation

	Valence	Arousal	Valence (normalization)	Arousal (normalization)
Sad	5.268984	−3.456652	0.259429587	0
Surprise	−5.86238	−0.467735	0	0.710211035
Enjoy	37.04469	−0.219423	1	0.769213643
Disgust	−0.30834	0.5307271	0.129443438	0.94746035
Fear	0.886882	0.5555072	0.157299522	0.953348456
Anger	−0.37264	0.3482074	0.127944947	0.904090959

Using Welch's method, we extracted the mean PSD feature from alpha band and beta band of frontal EEG as showed in Table 3. Then, we calculated the value of valence and arousal (Table 4). After that, we represented these emotional states in the valence—arousal two-dimensional space. The results show that the calculated valence and arousal values, when being mapped on the circumplex model, fit with the model. Five out of six evoked emotions were mapped in the right corresponding quarter region in the original model. The difference is only apparent with respect to the emotion of surprise (Fig. 3).

For facial expression data, we performed a blind analysis of facial expressions with the help of an expert in this field. The expert received 12 recorded videos of participant expression (6 videos for each participant) without sound and did not know the sequence in which emotions were evoked. The expert was asked to rate and point out which is the main emotion of participant' facial expression in each video. The results showed that only disgust was transiently expressed in the facial expressions of the participant, while for emotions of sadness, surprise, enjoyment, fear and anger, participant only expressed her faces in neutral state (Table 5).

5 Discussion and Conclusion

The results presented in Chap. 4 strongly support the emergent opinion that, facial expressions cannot be the reliable reflection of a person's innermost emotions they were once thought to be. The results of this study are not alone as there are several

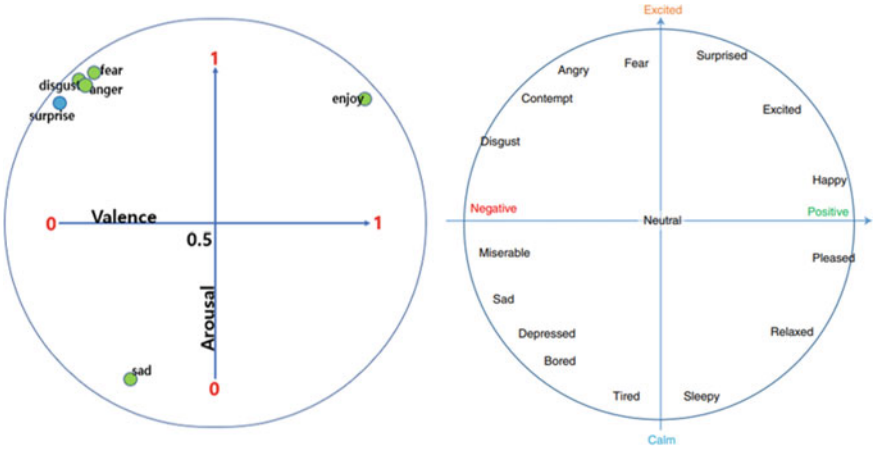


Fig. 3 Mapping the six emotional states in the circumplex model (left) The six emotional states identified by valence and arousal values in Table 4; (right) Russell's original model

Table 5 Comparison between emotion reported by the participant. and facial expression rated by the expert

Emotion reported by the participant	Facial expression rated by the expert	Emotion reported by the participant	Facial expression rated by the expert
Sad	Neutral	Disgust	Neutral, slightly Disgust
Surprise	Neutral	Fear	Neutral
Enjoy	Neutral	Anger	Neutral

other studies that agree with us, such as Barrett et al. (2019), Chen et al. (2018), Crivelli et al. (2017), Benitez-Quiroz et al. (2018) [3, 19, 20, 45].

This opinion poses a huge challenge for AI technology companies that are working on developing emotion recognition systems. Based on experimental results, relying only on facial expressions to identify emotions has been shown to be unilateral. Therefore, it is necessary to consider other expressions in gestures, behaviors, as well as biological indicators to be able to identify emotions more accurately [15].

While facial expressions did not show sufficient reliability, our study provides evidence that EEG signaling is a reliable tool for emotion detection and measurement. Only with PSD feature extraction, we have succeeded in modeling emotions on two-dimensional model of valence and arousal. This promises that EEG signal will be widely used in emotion recognition systems as input data in the future. Also, this simplicity, efficiency and reliability will help neuro-marketing and neuro-IS researchers in conducting their causal studies, especially when studying emotions as a mediating variable or a dependent variable.

Our study has limitations that need to be overcome in the future. Firstly, the data sample is small and lacks diversity. In the future, we will try to conduct experiments on more subjects, in diverse cultures and in more diverse experimental environments.

Second, the circumplex model is limited in clearly categorizing emotions. Valence and arousal measurements led to the clustering of emotions presented on the circumplex model. Thus, the circumplex model cannot be relied on to classify emotions. In the future, we will conduct research on models with more than two dimensions to be able to better separate emotional states.

Third, this study only used Ekman's 6 basic emotions in experimental design. However, recent studies have almost found common opinion that Ekman's six basic emotions are not enough to generalize human emotions. For example, Cowen and Keltner [41] have captured 27 distinct categories of emotion bridged by continuous gradients and developed the semantic space theory of emotion which states that emotion is high dimensional. These new findings will form the background for us to develop future studies.

Fourth, looking at our results, when trying to map six emotions onto the circumplex model, only the surprise emotion was in a position that doesn't match the original circumplex model. In the original model, the emotion of surprise has a positive tendency (positive valence value), but in our experiment, the emotion of surprise has a negative valence value, i.e., has a negative tendency. This leads to an interesting question that we want to find answers to in the future: When is the emotion of surprise positive and when is it negative?

Fifth, our study did not consider the neutral state. In our opinion, neutral means no emotion, so it is difficult to elicit neutrality when conducting experiments. We assume that a person can only achieve "neutral emotion" when his brain's activity is reduced to the maximum, that is, only when he has fallen asleep, or has been senseless. The validity of this assumption is currently untested, and the study of neutral emotion is an interesting topic that we will focus on in the future.

Sixth, not as simple as collecting face data, collecting EEG signals is complicated for most normal people under normal conditions. It requires specialized equipment and software along with meticulous installation. The inconvenience that it brings to the user is also a big obstacle for such devices to be used in everyday life, or even in laboratories that do not specialize in brain research. Therefore, in the future, studies and inventions for reducing the size and increasing the convenience of EEG measuring devices are essential. How to integrate EEG measurement sensors into common wearable devices, such as virtual reality headsets, is not a bad idea to work on in the future.

References

1. Gannouni, S., Aledaily, A., Belwafi, K., Aboalsamh, H.: Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. Sci. Rep. **11**(1), 1–17 (2021)
2. Meiselman, H.L. (Ed.): Emotion Measurement. Woodhead publishing (2016)
3. Barrett, L.F., Lewis, M., Haviland-Jones, J.M. (Eds.): Handbook of Emotions. Guilford Publications (2016)
4. Panksepp, J: Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford university press (2004)
5. Damasio, A.R.: Emotion in the perspective of an integrated nervous system. Brain Res. Rev. **26**(2–3), 83–86 (1998)
6. Ekman, P.E., Davidson, R.J: The Nature of Emotion: Fundamental Questions. Oxford University Press (1994)
7. Cabanac, M.: What is emotion? Behav. Proc. **60**(2), 69–83 (2002)
8. Schacter, D., Gilbert, D., Wegner, D., Hood, B.M.: Psychology: European Edition. Macmillan International Higher Education (2011)
9. Pinker, S.: How the Mind Works, vol. 524. Norton, New York (1997)
10. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
11. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**, 259–275 (2003)
12. Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Wang, X.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans. Multimedia **12**(7), 682–691 (2010)
13. Ekman, P., Sorenson, E.R., Friesen, W.V.: Pan-cultural elements in facial displays of emotion. Science **164**(3875), 86–88 (1969)
14. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
15. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol. Sci. Public Interes. **20**(1), 1–68 (2019)
16. Heaven, D.: Why faces don't always tell the truth about feelings. Nature **578**(7796), 502–505 (2020)
17. Ekman, P.: Facial expression and emotion. Am. Psychol. **48**(4), 384 (1993)
18. Gu, S., Wang, F., Patel, N.P., Bourgeois, J.A., Huang, J.H.: A model for basic emotions using observations of behavior in drosophila. Front. Psychol. **781** (2019)
19. Chen, C., Crivelli, C., Garrod, O.G., Schyns, P.G., Fernández-Dols, J.M., Jack, R.E.: Distinct facial expressions represent pain and pleasure across cultures. Proc. Natl. Acad. Sci. **115**(43), E10013–E10021 (2018)
20. Crivelli, C., Russell, J.A., Jarillo, S., Fernández-Dols, J.M.: Recognizing spontaneous facial expressions of emotion in a small-scale society of Papua new Guinea. Emotion **17**(2), 337 (2017)
21. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. **17**(3), 715–734 (2005)
22. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nat. Mach. Intell. **3**(1), 42–50 (2021)
23. Russell, J.A., Carroll, J.M.: On the bipolarity of positive and negative affect. Psychol. Bull. **125**(1), 3 (1999)
24. Watson, D., Clark, L.A.: Affects separable and inseparable: on the hierarchical arrangement of the negative affects. J. Pers. Soc. Psychol. **62**(3), 489 (1992)

25. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161 (1980)
26. Rubin, D.C., Talarico, J.M.: A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words. Memory **17**(8), 802–808 (2009)
27. Remington, N.A., Fabrigar, L.R., Visser, P.S.: Reexamining the circumplex model of affect. J. Pers. Soc. Psychol. **79**(2), 286 (2000)
28. Hamann, S.: Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends Cogn. Sci. **16**(9), 458–466 (2012)
29. Abhang, P.A., Gawali, B.W., Mehrotra, S.C.: Technological basics of EEG recording and operation of apparatus. In: Introduction to EEG-and Speech-Based Emotion Recognition, pp. 19−50 (2016)
30. Jenke, R., Peer, A., Buss, M.: Feature extraction and selection for emotion recognition from EEG. IEEE Trans. Affect. Comput. **5**(3), 327–339 (2014)
31. Welch, P.: The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967)
32. Bos, D.O.: EEG-based emotion recognition. Influ. Vis. Audit. Stimuli **56**(3), 1–17 (2006)
33. Schubring, D., Schupp, H.T.: Emotion and brain oscillations: high arousal is associated with decreases in alpha-and lower beta-band power. Cereb. Cortex **31**(3), 1597–1608 (2021)
34. Hwang, S., Jebelli, H., Choi, B., Choi, M., Lee, S.: Measuring workers' emotional state during construction tasks using wearable EEG. J. Constr. Eng. Manag. **144**(7), 04018050 (2018)
35. Rusinov, V.S.: Electrophysiology of the Central Nervous System. Springer Science & Business Media, (2012)
36. Blaiech, H., Neji, M., Wali, A., Alimi, A.M.: Emotion recognition by analysis of EEG signals. In: 13th International Conference on Hybrid Intelligent Systems (HIS 2013), pp. 312−318. IEEE (2013, December)
37. Ramirez, R., Vamvakousis, Z.: Detecting emotion from EEG signals using the emotive epoc device. In: International Conference on Brain Informatics, pp. 175−184. Springer, Berlin, Heidelberg (2012)
38. Kim, H.S., Lee, J.H.: Neuro-scientific approach to fashion visual merchandising-comparison of brain activation to positive/negative VM in fashion store using fNIRS. J. Korean Soc. Cloth. Text. **41**(2), 254–265 (2017)
39. Benitez, D.S., Toscano, S., Silva, A.: On the use of the Emotiv EPOC neuroheadset as a low cost alternative for EEG signal acquisition. In: 2016 IEEE Colombian Conference on Communications and Computing (COLCOM), IEEE, 1−6 April 2016
40. General Assembly of the World Medical Association: World medical association declaration of helsinki: ethical principles for medical research involving human subjects. J. Am. Coll. E Dent. **81**(3), 14–18 (2014)
41. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. **114**(38), E7900–E7909 (2017)
42. Available: https://www.emotiv.com/
43. Brunner, C., Delorme, A., Makeig, S.: Eeglaban open source matlab toolbox for electrophysiological research. Biomed. Eng./Biomed. Tech. **58**(SI-1-Track-G), 000010151520134182 (2013)
44. Winkler, I., Haufe, S., Tangermann, M.: Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behav. Brain Funct. **7**(1), 1–15 (2011)
45. Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M.: Facial color is an efficient mechanism to visually transmit emotion. Proc. Natl. Acad. Sci. **115**(14), 3581–3586 (2018)

NSGA-II-AMO: A Faster Genetic Algorithm for QWSCP

Zehui Feng, Bei Wang, Mingjian Chen, and Qi Chen

Abstract The task of the QoS-driven web service composition problem (QWSCP) is to find the optimal combination of atomic web services to meet the high demand for QoS parameters. However, different scenarios have various sensitivities to each parameter, researchers prefer algorithms that solve the problem to output non-dominated solution sets. NSGA-II, a genetic algorithm, has become one of the most contacted algorithms for solving the QWSCP problem due to its ability to search for non-dominated solution sets. However, NSGA-II algorithm still suffers from the challenge of slow convergence and the tendency to fall into local optimum solutions. In adjusting the NSGA-II mutation probabilities, it was found that the setting of the mutation probabilities affects the effectiveness of the algorithm, and that the optimal mutation probabilities for different loci are related to the distribution of data for that locus. Therefore, this paper proposes an improved NSGA-II-AMO algorithm, which uses an adaptive mutation operator to complete the mutation process when generating children in the NSGA-II algorithm, so that the mutation probability adaptively changes with different distributions of data, effectively improving the search speed at the beginning of population iteration; at the same time, in order to avoid the problem of oscillatory search at a later stage, AMO reduces the mutation probability as the population converges, preventing the re-introduction of inferior genes into the

Z. Feng
Polytechnic Institude, Zhejiang University, Hangzhou, China
e-mail: zju_fengzehui@163.com

B. Wang · M. Chen · Q. Chen (✉)
College of Computer Science and Technology, Zhejiang University, Hangzhou, China
e-mail: zju_chenqi@163.com

B. Wang
e-mail: wangbei@zju.edu.cn

M. Chen
e-mail: chenmj01@163.com

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1_15

population. We compare the NSGA-II-AMO algorithm cross-sectionally with the PSO algorithm, the NSGA-II algorithm and the ABC algorithm, and the adaptive mutation operator achieves improved convergence with guaranteed distributivity.

Keywords QoS-driven · Service composition · NSGA-II · Adaptive mutation operator

1 Introduction

With the rapid growth of the Internet and the increasing complexity of web service requirements, service composition has become mainstream [1, 2]. How to select and combine a cluster of atomized services with high throughput, high availability, high reliability, low cost and low latency among the huge number of atomized services has become a hot issue of concern for the academic community [3]. An excellent selection, combination and filtering algorithm, mechanism, system and scheme can greatly assist in the rapid construction of services to achieve a portfolio of services with high Quality of Service (QoS) values. The task of QWSCP can be described as selecting a set of atomic services from a given set of optional atomicity services, such that the combined set of atomic services fulfils the functional requirements of the user while also having QoS parameters that satisfy the non-functional requirements of the user.

A large body of works have been proposed to solve the service composition problem, mainly in the form of linear integer programming model (LIPM) methods, Metaheuristic-based methods and Multi-objective optimization methods. The LIPM approach treats the service composition problem as a linear integer programming model that allows an optimal composition to be found [4, 5]. In the metaheuristic approach, the service composition problem is then treated as a single-objective optimization problem (SOOP) [6, 7], which returns an optimal or near-optimal solution. However, the evaluation of service composition solutions is personalized and the importance of each QoS parameter is different in different business scenarios and software styles. For example, high real-time and low latency in some scenarios, and high reliability and availability in others. Therefore the ability of the above two service composition solutions with a single output solution to meet the mutually exclusive needs of the users is highly dependent on the expertise of the decision maker and the design of the QoS parameter weights. For these reasons, more and more researchers are beginning to view service composition as a multi-objective optimization problem.

In this paper, the NSGA-II-AMO algorithm is proposed to solve the service composition problem, and three adaptive mutation operators are proposed according to the data distribution of each gene locus, bringing an improvement in the convergence effect while ensuring the distributional effect of the genetic algorithm.

2 Related Work

In recent years, the number of service composition has grown rapidly to meet the needs of users, and how to quickly and efficiently achieve a reasonable selection of combined services from optional atomic service composition has become the focus of many scholars' research. Yin et al. [8] proposed a service composition optimization model that considers eight QoS variables, such as response time, quality of service, cost, availability, and concordance, and uses grey correlation analysis to obtain the correlation between these QoS attributes so that the QoS metrics of the combined solution can be improved. Yu [9] proposed a computational model that can calculate skylines according to a service composition model, enabling service users to optimally access a service set in the form of an integrated service package, first by a one-pass algorithm determined entirely by a single service skyline, and then using a double asymptotic algorithm A comparison is made to give the optimal skyline composition of services.

However, as the above methods can only produce a single solution, more and more multi-objective optimization methods are being proposed to solve the QWSCP problem. The main advantage of multi-objective optimization methods is that this class of methods ultimately outputs a set of non-dominated solutions from which the user can select the non-dominated solution that best meets his or her needs. Among them, multi-objective genetic algorithms have been extensively studied to optimize multi-objective parameters simultaneously to solve service composition problems [10]. For example, Niu et al. [11] modelled the process of constructing a service composition as a constrained interval number multi-objective optimization problem and proposed a decomposition-based uncertain multi-objective evolutionary algorithm , which was used to build a mathematical model to calculate the optimal service portfolio. Meanwhile, researchers have proposed some multi-objective versions of meta-heuristic single-objective optimization algorithms to solve the service composition problem. For example, Sadouki et al. [12] proposed a multi-objective version of an elephant grazing optimization algorithm; Wang [13] et al. proposed a QoS-aware service-selection-specific algorithm based on the artificial bee colony algorithm (ABC) in the discrete space for efficient local search. However, the above multi-objective optimization algorithm requires the decision maker to set the variance probability, too large a variance probability will lead to oscillation in the final stage of the search, too small a variance probability will cause the search process to stall prematurely. The convergence effect of the above multi-objective optimization method is extremely demanding in terms of matching the variance probability to the data samples, which makes it difficult to meet the convergence requirements of the user when performing service compositions.

3 Problem Formulation

The task of QoS-driven web service composition problem is to select specific services to be filled into each node of the combined service, so that the combined service has the best possible QoS parameters. When viewed as a multi-objective optimization problem, this problem can be described as the following 5-tuple $(P, S, \Omega, F(x), R)$ with the components shown below.

- P is the concrete process of the combined service, $(p_1, p_2, \cdots, p_n)$ describes the abstract atomic service types of n nodes which the combined service consists of.
- S is the set of optional subservices, each with an abstract service type and QoS parameters.
- Ω is the space where the decision vector $x = (x_1, x_2, \cdots, x_n)$ located in. The vector x represents the subservices $(s_1, s_2, \cdots, s_n)$ selected in turn for combining each node of the combined service, where $s_1, s_2, \cdots, s_n \in S$. Also, s_i and p_i have the same abstract service type, which means they can perform similar functions.
- $F(x) = (f_1(x), f_2(x), ..., f_m(x))$, is the specific optimization objective, where m is the number of QoS parameters and $x \in \Omega$. In this problem, the optimization objective can be expressed as max $F(x)$ by operations such as data normalization. In our research, four QoS parameters were selected: availability, reliability, response time and throughput, so $m = 4$.
- R is the set of results, $R = (x_1, x_2, \cdots, x_w)$, where $x_1, x_2, \cdots, x_w \in \Omega$ and are non-dominated relations with each other. w is the number of individuals in this non-dominated solution.

The key of QoS-driven web service composition problem is to make the F of the members in R as big as possible, given that P, S, Ω and $F(x)$ are determined. As stated above, different users have different sensitivities and needs for different $F(x)$. In most cases, however, it is difficult to find a x_i, such that all four dimensions of $F(x_i)$ are optimal, but rather only a set of x for which $F(x)$ has different dominance, so the problem ultimately requires that the output set of results R is a non-dominated set of x.

4 Our Solution

We propose the NSGA-II-AMO algorithm to solve the QoS-driven web service composition problem. NSGA-II-AMO is an improvement on the NSGA-II algorithm. NSGA-II improves the fitness of a population by constructing a population full of competition, through elimination, crossover and mutation over and over again. In applying the NSGA-II algorithm to solve QoS-driven web service composition problem, each decision vector $\boldsymbol{x}$ is considered as an individual, each part on vector $\boldsymbol{x}$ is considered as a genetic loci of the individual, and the set of multiple decision vectors $\boldsymbol{x}$ is considered as a population.

The main flow of the NSGA-II-AMO algorithm is as follows:

Step 1: Primary population production

Set the number of primary populations to N. This step randomly generates N decision variables $(x_1, x_2, \cdots, x_N)$ as the incipient parent population P_0, where $x_1, x_2, \cdots, x_N \in \Omega$.

Step 2: Generation of children

Select two individuals as parents from the parent population Pt and generation a new child by parental crossover. Then repeat until N children individuals are produced to form the children population Q_t. At the same time, there is a probability of variation during the generation of individuals, expanding the search area. It is important to set the probability of variation in the generation of children. A large probability of variation can lead to the replacement of introduced superior genes by inferior genes, causing oscillatory search in the late convergence period; a too small probability of variation can make it difficult to introduce optional superior genes from outside the population and stall the search process prematurely. When the probability of variation is the same for all loci, this value is too large for some loci and too small for another, making it difficult to achieve good convergence. Therefore, this paper proposes three adaptive mutation operators (AMO) that calculate the probability of variation separately according to the data distribution of each locus.

- **AMO1:** We first design an adaptive mutation operator based on the number of selectable genes at each locus, which calculates the probability of mutation for each locus as in:

$$\lambda_1 = \exp\left(\frac{\text{num} - \min}{\max - \min} - 0.5\right)\lambda_0 \tag{1}$$

  where num is the number of genes available for the locus, max and min are the maximum and minimum values of the number of genes available for all loci, respectively, and λ_0 is the base variation probability, usually set to 0.05. In AMO1, the probability of mutation is simply related to the number of available genes at each locus, with the larger the number, the higher the probability of mutation.
- **AMO2:** We designed AMO2 by considering the effect of the degree of data dispersion, which calculates the probability of mutation for each locus as in:

$$\lambda_2 = \sum_{i=1}^{n} \frac{s_i}{\overline{QoS_i}} \tag{2}$$

  where n is the number of QoS parameters, 4 in this experiment; s_i is the standard deviation of the ith parameter for that locus and $\overline{QoS_i}$ is the mean. λ_2 reflects the data dispersion for each locus, the greater the data dispersion, the greater the probability of variation.
- **AMO3:** As mentioned earlier, when using a fixed variation probability, too large a variation probability at a later stage can cause oscillatory search problems, so we designed AMO3 with a negative correlation between the variation probability and

the number of Generations to solve this problem, which calculates the probability of mutation for each locus as in:

$$\lambda_3 = (\exp(-gen/40) + 0.5)\lambda_2 \tag{3}$$

where gen is the number of generations. AMO3 has a progressively lower probability of variation as the number of population generations increases, based on AMO2 considering the degree of dispersion.

Step 3: Combine P_t and Q_t into a new population R_t
NSGA-II adopts a retention strategy for the parent, and the new generation population Rt consists of the parent population Pt and the offspring population Q_t combined, i.e. $R_t = P_t \cup Q_t$.
Step 4: Sorting
An undominated sort is performed on R_t, and a further crowding sort is performed on individuals of the same undominated rank. These two kinds of sorting ensure the richness of the population, allowing the search for the optimal solution to remain in different directions.

Since the results of the ranking are seriously affected by the different magnitudes of the different QoS in the process of crowding sorting, the data need to be normalized. In our paper, we use a min-max normalization as in:

$$QoS_+ = \frac{QoS - \min}{\max - \min} \tag{4}$$

$$QoS_- = 1 - \frac{QoS - \min}{\max - \min} \tag{5}$$

where QoS_+ and QoS_- are the results after data normalization, QoS is the original value before normalization, max is the maximum value of the parameter and min is the minimum value of the parameter. Equation (4) is used for larger and better QoS parameters, such as reliability, availability and throughput, and Equation (5) is used for smaller and better QoS parameters, such as response time. After the above data normalization operation, all four QoS parameters were converted to parameters whose values were distributed within [0, 1] and the larger the better, eliminating the influence of the different parameter's magnitudes on the ranking.

When sorting, we need to calculate the QoS parameters of the services after the combination by using the QoS parameters of the atomic services. It is worth noting that the calculation of this step varies with the topology of the services. The topology of the services can be divided into four main types: sequence, cyclic, parallel and branch, whose topology and calculation are shown in Fig. 1 and Table 1.
Step 5: Select the first half of the individuals in R_t *to form a new parent population* P_{t+1}
The algorithm will repeat step 2–step 5, increasing the fitness of the population through continuous filtering. When the end condition is reached, (e.g. the fitness

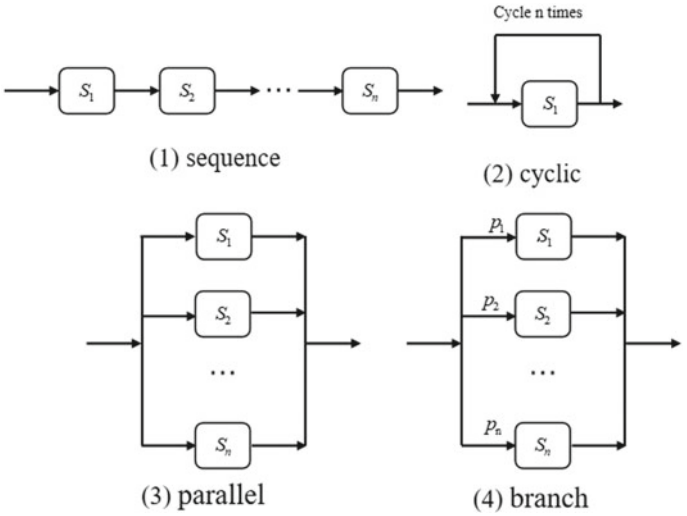


Fig. 1 Example of four topologies

Table 1 QoS calculation for four topologies

	Sequence	Cyclic	Parallel	Branch
Reliability R	$\prod_{i=1}^{n} R_i$	R^n	$\prod_{i=1}^{n} R_i$	$\sum_{i=1}^{n} p_i R_i$
Availability A	$\prod_{i=1}^{n} A_i$	A^n	$\prod_{i=1}^{n} A_i$	$\sum_{i=1}^{n} p_i A_i$
Throughtout T	$\text{Min}_{i=1}^{n} T_n$	T	$\text{Min}_{i=1}^{n} T_n$	$\sum_{i=1}^{n} p_i T_i$
Response time RT	$\sum_{i=1}^{n} RT_i$	$n * RT$	$\text{Max}_{i=1}^{n} RT_n$	$\sum_{i=1}^{n} p_i RT_i$

has converged or the number of iterations has been reached, etc.), the final set of non-dominated solutions of the filtered population is output.

5 Algorithm Implementation and Results

5.1 Dataset

The dataset used in this paper is the QWS Dataset Version 2.0, which collects information on 2507 web services, containing eleven attributes for each service and a functional classification. The four QoS parameters of availability, reliability, response

time and throughput for each service are used in this study, while services with the same functional classification are considered as atomic services with similar functions.

5.2 Evaluation Indicators

This paper uses 2 metrics recognized in the field for the evaluation of multi-objective optimization problems, a measure of their convergence and a measure of their population diversity [14].

- Convergence indicator
  The convergence indicators are evaluated as in:

$$r = \sum_{i=1}^{n} d_i / n \tag{6}$$

  where n is the number of solution vectors in the final output non-dominated solution set; d_i is the Euclidean distance between each solution vector in the obtained non-dominated solution set and the nearest member of the true Pareto front (measured in the optimization objective parameter space after data normalization). This indicator mainly reflects the convergence effect of the multi-objective optimization algorithm. The lower the value of r, the better the population convergence.
- Distributional Indicator
  The distributional indicators are evaluated as in:

$$d = \frac{\sum_{i=1}^{n-1} d_i}{n-1} \tag{7}$$

$$\Delta = \frac{d_f + d_l + \sum_{i=1}^{n-1} \left| d_i - \bar{d} \right|}{d_f + d_l + (n-1)d} \tag{8}$$

  where d_i is the Euclidean distance between the ith individual and the $(i+1)$st individual in the final output set of non-dominated solutions, d_f and d_l represent the distance between the extreme value solution of the true Pareto frontier and the solution at the boundary of the output set, and d is the average of d_i. As mentioned earlier, the importance attached to different parameters varies from scenario to scenario, users tend to expect a better distribution of the solution set obtained. The smaller the value of Δ, the better the distribution.

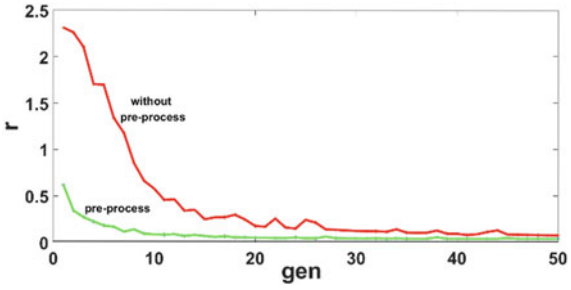


Fig. 2 Data pre-processing effect

5.3 Data Pre-process

In the set S of atomic services, the number of atomic services providing the same function is quite large. If there exist atomic services s_i and s_j, satisfying $Q_r(s_i) \leq Q_r(s_j)$, $Q_a(s_i) \leq Q_a(s_j)$, $Q_t(s_i) \leq Q_t(s_j)$, $Q_{rt}(s_i) \leq Q_{rt}(s_j)$, it means that s_i does not exceed s_j in all four QoS parameters, in other words, s_i is an inferior gene at that locus, and selecting s_i will result in an individual with inferior QoS parameters to selecting s_j of similar individuals and should therefore be eliminated directly. Therefore, during the data pre-processing process, the skyline algorithm was used to eliminate inferior genes like s_i. The above data pre-processing operation can greatly improve the speed of population convergence.

Figure 2 shows the effect of data pre-processing on the speed of population convergence. It can be seen that when no data pre-processing is performed, the initial population convergence index is very poor due to the large number of "inferior" genes incorporated into the initial population generation, and it takes fifty iterations before the initial pre-processed population converges.

5.4 Selection of the Initial Population Size N

This section verifies the effect of the initial number of populations N on the convergence effect. To ensure the fairness of the experiment, the experiment tests the metrics of the output non-dominated solution set after the same running conditions and running time. The setting of the initial number of populations, N, must have an effect on the convergence effect. When N is too large, the number of computations will increase in non-polynomial time, resulting in fewer iterations and slower convergence in the same time; when N is too small, a large number of initial populations cannot enter the initial population, resulting in low genetic richness of the population itself. Therefore, in this section set the N values to 60, 80, 100, 120, 140. The experimental results are shown in Table 2.

As can be seen from Table 2, the distributional indicator gradually becomes better as the population size increases, while the convergence indicator becomes better in

Table 2 Convergence for different values of N

	PSO	ABC	NSGA-II	NSGA-II-AMO3(ours)
r	0.0455	***0.0356***	0.0397	0.0367
Δ	0.3428	0.2787	0.2531	***0.2501***

Table 3 Results for different AMO

p mutation operator	Fixed	AMO1	AMO2	AMO3
r	0.0397	0.0391	0.0382	**0.0367**
Δ	0.2531	0.2544	0.2511	**0.2501**

60–80 and worse in 80–140. It is also worth noting that changes in N values have a greater impact on r, with the difference between the best and worst reaching 8.43%, while the impact on is smaller, with the difference between the best and worst being only 1.42%. Therefore, the initial population size N was set to 80 in the subsequent experiments.

5.5 Selection of AMO

We propose three adaptive mutation operators: AMO1 is a consideration of the number of optional genes, and the probability of variation increases with increasing number; AMO2 is a consideration of the data dispersion of optional genes, and the greater the data dispersion, the higher the probability of variation; AMO3 adds a consideration of the current generation to AMO2, and the probability of variation gradually decreases with increasing population generation. Similarly, we conduct experiments on the QWS dataset, which will use the NSGA-II algorithm of fixed mutation value and three AMOs run for the same time under the same conditions, for which the r and Δ values of the non-dominated set (Table 3).

It can be seen that all three variational operators achieve an improvement in r values, especially AMO2 and AMO3, which improve in both metrics as they take into account the degree of data dispersion and can mitigate the effects of local optimal solutions. On the comparison between AMO2 and AMO3, it can be seen that AMO3 works better. AMO3 is 3.93% ahead of AMO2 in r value and 0.40% ahead in value. From the calculation of the variation probabilities of amo2 and amo3, it can be seen that the variation probability of AMO3 is higher than that of AMO2 before 27 generations, while the opposite is true after 27 generations. In other words, in the early stages of convergence, when a large number of "good" genes have not yet entered the population, AMO3 has a higher probability of mutation and can introduce these genes faster; while after a period of convergence, AMO3 has a lower probability

Table 4 Results for different algorithms

N	60	80	100	120	140
r	0.0401	***0.0397***	0.0398	0.0422	0.0427
Δ	0.2535	0.2531	0.2511	0.2517	***0.2499***

of mutation and can effectively prevent shock search or mutation to produce "bad" genes. Therefore, AMO3 has a better convergence effect than AMO2.

5.6 Comparison with Other Algorithms

The NSGA-II algorithm, the Particle Swarm optimization (PSO) algorithm and the Artificial bee colony(ABC) algorithm were selected as comparison algorithms and tested in the same way to compare their r and Δ values.

As can be seen from Table 4, thanks to the application of AMO3, the Δ value of our method compared to NSGA-II improved from 0.2531 to 0.2501, an improvement of 1.19%, and the r value improved from 0.0397 to 0.0367, an improvement of 7.56%. The main reason for this improvement is that our adaptive variation operator takes into account the effects of data distribution and population generation, and the variation probabilities are changed adaptively so that the variation probabilities at each locus match the current data distribution and iteration level more closely at all times, and therefore converge better during the iteration. Compared to the ABC algorithm, our algorithm lags behind in r value by 3.00%, but improves in Δ from 0.2787 to 0.2501, an improvement of 10.26%. In other words, our algorithm suffers from a relatively small lag in convergence relative to the ABC algorithm, but has a large advantage in distributivity. The main reason for this effect is that our algorithm devotes a significant amount of computational time to ensuring distributivity, including the non-dominated sorting and congestion sorting processes at each iteration. As we stated earlier, users of the combined service have different needs for the four parameters of the service, so a better population richness helps to match the needs of a more diverse set of users.

6 Conclusion

In this paper, we propose the NSGA-II-AMO algorithm, which adds an adaptive variation operator to the NSGA-II algorithm, allowing the variation probability to change adaptively and dynamically with the data distribution and the number of population iterations. Thanks to the introduction of AMO, our algorithm achieves a 7.56% improvement in convergence compared to the original NSGA-II algorithm

on the QWS dataset, with a 1.19% improvement in distributivity, and performs well against other algorithms. Overall, our algorithm achieves an improvement in convergence speed while maintaining distributivity.

Acknowledgements This research is funded by the National Key Research and Development Program of China (Grant No. 2018YFB2101200) and ZJU-Guangxi Province Collaborative Project (Grant No. ZD20302004).

References

1. Yuan, Y., Zhang, W.S., Zhang, X.G.: A context-aware self-adaptation approach for web service composition. In: 2018 3rd International Conference on Information Systems Engineering (ICISE), pp. 33–38 (2018)
2. Gustavo, A.: Review: Perspectives on web services—applying SOAP, WSDL and UDDI to Real-World Projects. IEEE Internet Comput. **47**, 505–505 (2018)
3. Jatoth, C., Gangadharan, G., Buyya, R.: Computational intelligence based QoS-aware Web service composition: A systematic literature review. IEEE Trans. Serv. Comput. **10**, 475–492 (2015)
4. Ghobaei-Arani, M., Souri, A.: LP-WSC: A linear programming approach for Web service composition in geographically distributed cloud environments. J. Supercomput. **75**, 2603–2628 (2019)
5. Alrifai, M., Risse, T., Nejdl, W.: A hybrid approach for efficient web service composition with end-to-end QoS constraints. ACM Trans. Web **6**(2), 7 (2012)
6. Dahan, F., El Hindi, K., Ghoneim, A.: Enhanced artificial bee colony algorithm for QoS-aware Web service selection problem. Computing **99**, 507–517 (2017)
7. Huo, Y., Zhuang, Y., Gu, J., Ni, S., Xue, Y.: Discrete gbest-guided artificial bee colony algorithm for cloud service composition. Appl. Intell. **42**, 661–678 (2015); (Int. J. Artif. Intell. Neural Netw. Complex Probl. Solving Technol.)
8. Yin, C., Zhang, Y., Zhong, T.: Optimization model of cloud manufacturing service resource combination for new product development. Comput. Integr. Manuf. Syst. **18**, 1368–1378 (2012)
9. Yu, Q., Bouguettaya, A.: Computing service Skylines over sets of services. In: Proceedings of the IEEE International Conference Web Services. IEEE Computer Society, pp. 481–488. Washington DC (2017)
10. Cremene, M., Suciu, M., Pallez, D., et al.: Comparative analysis of multi-objective evolutionary algorithms for QoS-aware web service composition. Appl. Soft Comput. **39**, 124–139 (2016)
11. Niu, S., Zou, G., Gan, Y., et al.: Towards uncertain QoS-aware service composition via multi-objective optimization. In: IEEE 24th International Conference Web Service, 25–30 June 2017, pp. 894–897. IEEE, Honolulu, HI, USA (2017)
12. Sadouki, S.C., Tari, A.: Multi-objective and discrete elephants herding optimization algorithm for QoS aware web service composition. RAIRO-Oper. Res. **53**, 445–459 (2019)
13. Wang, X.Z., Wang, Z.J., & Xu, X.F.: An improved artificial bee colony approach to QoS-aware service selection. In: Proceedings of the 20th IEEE International Conference Web Service. IEEE Computer Society, pp. 395–402. Washington DC (2013)
14. Seghir, F.: FDMOABC: Fuzzy discrete multi-objective artificial bee colony approach for solving the non-deterministic QoS-driven web service composition problem. Expert Syst. Appl. **167**, 114413 (2021)

Author Index

R. Lee (ed.), *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Studies in Computational Intelligence 1074,
https://doi.org/10.1007/978-3-031-19604-1